DOCTOR OF PHILOSOPHY

Approaches to understanding diversity in rubber and carotenoid synthesis in Hevea brasiliensis latex

Bahari, Azlina

Award date:
2019

Approaches to understanding diversity in rubber and carotenoid synthesis in *Hevea brasiliensis* latex

Azlina Bahari

A thesis presented for the degree of Doctor of Philosophy

School of Life Sciences

University of Dundee

Information & Computational Sciences group

The James Hutton Institute

January 2019

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgment, the work presented is entirely my own.

Azlina Bahari 18th December 2018

David F. Marshall 18th December 2018
PhD supervisor
Information and Computational Sciences Group
The James Hutton Institute
University of Dundee

Table of contents

List of tables

Chapter 1

Chapter 2

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 5

Chapter 6

List of abbreviations used

AACT	Acetoacetyl coenzyme A
AU	Arbitrary unit
BPC	Base peak chromatogram
cDNA	Complement deoxyribonucleic acid
CID	collision ion dissociation
CPS	Count per second
CPT	*cis*-prenyl transferase
DAD	Diode array detector
DMAPP	Dimethyl allyl diphosphate
DNA	Deoxyribonucleic acid
DRC	Dry rubber content
DXS	Deoxyxylulose phosphate synthase
EIC	Extracted ion chromatogram
ESI	Electrospray ionisation
EST	Expressed sequence tags
FDR	False detection rate
FPP	Farnesyl diphosphate
FW	Frey-Wyssling
GC	Gas chromatography
GPP	Geranyl diphosphate
GPS	Geranyl diphosphate synthase
GWAS	Genome-wide association study
HILIC	Hydrophilic liquid chromatography
HMG coA	Hydroxy methyl glutaryl coenzyme A

HMGR	Hydroxy methyl glutaryl coenzyme A reductase
HPLC	High-performance liquid chromatography
IPP	Isopentenyl diphosphate
Iso-seq	Isoform sequencing
KASP	Kompetitive Allele Specific PCR
m/z	mass-to-ion ratio
MDS	Multi-dimensional scale
MEP	Methylerythritol phosphate
MRM	Multi reaction monitoring
MS	Mass spectrometry
MVA	Mevalonate
NCBI	National Center for Biotechnology Information
NGS	Next generations sequencing
NMR	Nuclear magnetic resonance
PCR	polymerase chain reaction
PSY	Phytoene synthase
QTL	Quantitative trait loci
REF	Rubber elongation factor
REFSRPP	Rubber elongation factor/Small rubber particle protein gene family
RIN	RNA integrity number
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RT	Retention time
SNP	Single nucleotide polymorphism
SRA	Sequence Retrieval Archive

SRPP	Small rubber particle protein
TIC	Total ion chromatogram
TPM	Transcripts per million
UTR	Untranslated region

ACKNOWLEDGEMENTS

I wish to thank my supervisors, Dr David Marshall (The James Hutton Institute), Dr Mark Taylor (The James Hutton Institute), Dr Raymond Campbell (The James Hutton Institute), Dr Chow Keng See (Malaysian Rubber Board) for their guidance and advice through this course of study. In addition, a token of appreciation to Prof Claire Halpin (University of Dundee) for her time and guidance.

I am grateful for the technical assistance and advice from several staff from the James Hutton Institute: Dr Will Allwood with the metabolite work, Dr Pete Hedley and Mrs Jenny Morris for their help in constructing cDNA libraries; Dr Micha Bayer and Mr Paulo Flores for their input on transcriptome analysis; Mr Simon Pont, Brian Harrower and Mr Ralph Wilson for the potato materials; Ms Katrin Mackenzie for the phylogenetic tree analysis and Dr Kelly Houston for her help in KASP genotyping assay.

I would like to acknowledge the technical assistance of the Malaysian Rubber Board Staff (Ms Zainorlina Mohd Zainuddin, Ms Azlina Azharuddin, Ms Noraini, Mr Mohd Firdaus Razaleigh, Mr Monyrajan Venugal and Mr Mohd Fahmi) in collecting and processing the latex samples. I would like to acknowledge Ms Nurmi Rohayu Abdul Majid and Mr Zarawi Abdul Ghani for their kind help in providing the information about *Hevea* breeding and yield data. I really appreciate help by Mr Ahmad Khairul and Ms Noor Hiza in analysing dry rubber content. In addition, a token of appreciation for Ms Siti Shuhada Shuib for her *Hevea* transcriptome data sharing.

I wish to thank all my friends at the James Hutton Institute, colleagues at the Plant Science Department, University of Dundee and friends at the Malaysian Rubber Board. My heartful thanks to the Malaysian Rubber Board management for giving me the chance to pursue this study. Lastly, to my beloved mum, thank you for the continuous prayer and unconditional love. To my Kak Long, siblings, nieces and nephews - I would like to express my special gratitude and appreciation for their continuous support and encouragement during this course of study.

ABSTRACT

Hevea brasiliensis latex contains a large quantity of high molecular weight rubber and is thus the primary commercial source of natural rubber. Rubber and other non-rubber isoprenoids in *Hevea* latex are synthesised from isopentenyl diphosphate (IPP) generated from the cytoplasmic mevalonate (MVA) pathway and the plastidic methyl erythritol phosphate pathway (MEP). This study utilised two rubber tree clones (RRIM600 and PB235) that show visibly contrasting levels of yellow carotenoids for the measurement of latex isoprenoids (carotenoids, rubber and isoprenoid intermediates) and transcript levels of the genes involved in isoprenoid biosynthesis. Metabolite extraction and analysis showed that four major carotenoids namely lutein, zeaxanthin, α-carotene and β-carotene were consistently present in both RRIM600 and PB235 latex. β-carotene was found to be the major carotenoid, at 1.2 µg/g in PB235 and 0.8 µg/g in RRIM600 fresh latex samples. However, the analytical method developed to measure isoprenoid intermediates needed to be further optimised to increase extraction efficiency. To enable accurate measurement of transcript levels of key genes involved in the isoprenoid biosynthetic pathway, a set of reference transcripts was constructed by merging short-reads (RNA-seq) and long-reads (Iso-seq and full-length cDNA sequences) data from *Hevea brasiliensis*. This produced a comprehensive set of 193,997 transcript sequences with good level of coverage of predicted transcripts and highly conserved core plant genes. Not only did the reference transcriptome update the annotation of rubber gene models, additional transcript variants were also discovered. Manual curation of gene models for key steps associated with rubber and carotenoids resulted in a repertoire of 115 genes, with 151 corresponding transcript variants. Subsequently, differential expression analysis

on the basis of mapping RRIM600 and PB235 RNA-seq reads to the reference transcriptome revealed isoform-specific expression of genes for biosynthesis of carotenoids (*PSY* isoform 2), IPP (*AACT2* and *HMGR1*) and rubber (REFSRPP gene members). In addition, the levels of these genes correlated positively with the carotenoid and rubber content measurements from the same latex of PB235 and RRIM600 used for metabolite extraction. Finally, the utility of the reference transcript catalogue was demonstrated by the characterisation of the REFSRPP gene family, which is involved in rubber elongation steps. REFSRPP gene family showed a local expansion which appear to be unique to *Hevea.* A pilot study has demonstrated there is considerable diversity of the genomic region containing the duplicated REFSRPP genes.

Chapter 1

General Introduction

1.1. The uniqueness of rubber crop from the latex of *Hevea brasiliensis*

Hevea brasiliensis is indigenous to the Amazonian rainforest in Brazil. In 1876, a total of 74,000 seeds were collected from Tapajos and Madeira regions of the Central Amazon basin. Of these seeds, only twenty-two seeds were successfully germinated in the then Malaya, and these seedlings have become the basis of Malaysia's rubber industry. Since then, the systematic cultivation and continuous improvement of *Hevea brasiliensis* has generated export earnings up to £8.1 billion per annum for Malaysia (Figure 1.1.1). While *Hevea* cultivation thrives in South East Asia, commercial plantations were not as successful in its place of origin. The trees planted in South America never reached physical maturity, due to a fungal disease, known as South American Leaf Blight (Lieberei, 2007).

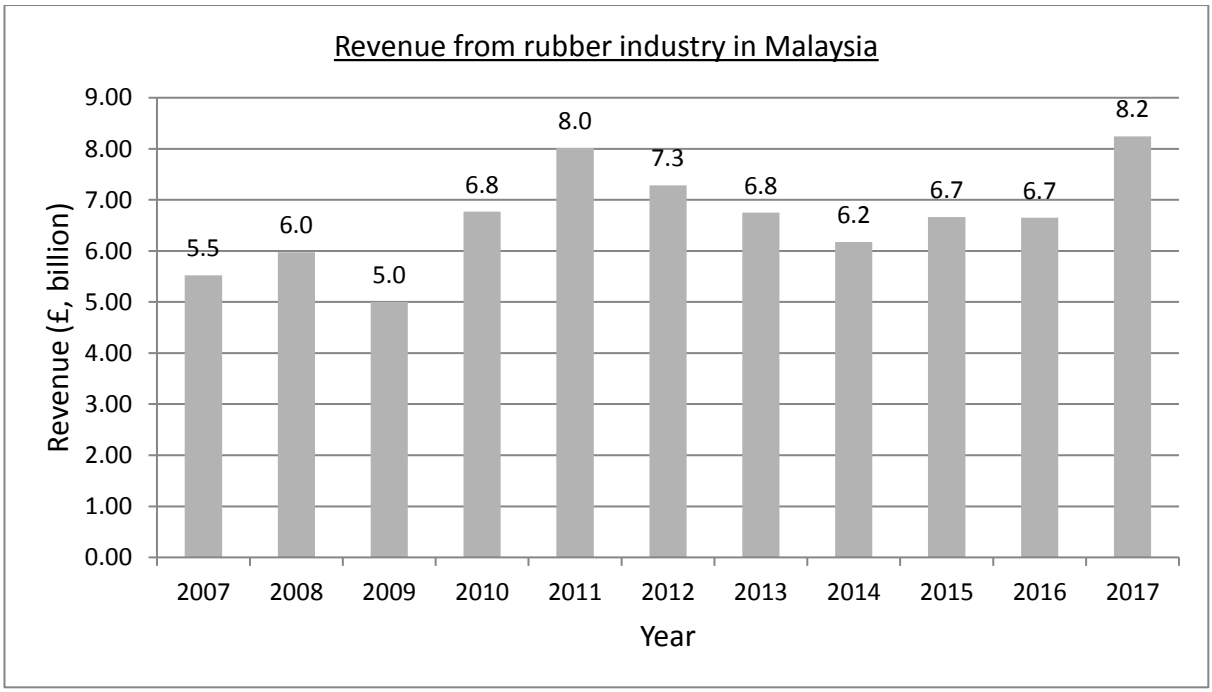


Figure 1.1.1.: The contribution of rubber industry to Malaysia's income. Source: Department of Statistics Malaysia (DOSM); Malaysian Timber Industry Board (MTIB)

Although latex is produced by some 12,500 plant species, only 1000 species contain rubber (Backhaus, 1985, Lewinsohn, 1991, Hunter, 1994). Of these, *Hevea* proved to the most commercially viable due to its latex containing

rubber with outstanding physical properties such as tensile strength and tear strength (Ho et al., 1975a, Soratana et al., 2017). Other rubber-producing plants, notably dandelion (*Taraxacum koksaghyz*) and guayule (*Parthenium argentatum*) are also used as rubber providers (van Beilen and Poirier, 2007). Yet, *Hevea* is the primary source of rubber due to three factors, namely i) the superior quality of its rubber; ii) quick latex regeneration; and iii) ease of harvesting. The main of features of rubber produced by the main rubber-producing plants are listed in Table 1.1.1.

The grading and pricing of latex is mostly dependent on dry rubber content (DRC). Although the DRC is dependent on genetic and environmental factors, *Hevea* latex DRC usually ranges from 30% to 50% (Eng et al., 2001, Ong, 2000), which is higher than the latex of other plant species.

Table 1.1.1: The main features of the rubber crop produced from the prominent rubber-producing plants, *Hevea,* dandelion (*Taraxacum koksaghyz*) and guayule (*Parthenium argentatum*)

	Hevea brasiliensis	***Taraxacum koksaghyz***	***Parthenium argentatum***
Latex location	Laticifer tissue (Gomez and Moir, 1979)	Root parenchyma (Schmidt et al., 2010)	Bark parenchyma (Hagel et al., 2008)
Rubber content	30-50% (Malaysian Rubber Board, 2009)	1-30% (van Beilen and Poirier, 2007)	3-15% (Cornish and Brichta, 2002)
Rubber molecular weight	$5x10^5$ – $2x10^6$ Dalton (Subramaniam, 1980)	$3x10^5$ Dalton (Buranov and Elmuradov, 2010)	$1x10^6$ – $1.6x10^6$ Dalton (Cornish and Brichta, 2002)
Harvesting approach	Non-destructive (tapping of the bark)	Uprooting the whole plant	Uprooting the whole plant

Hevea latex is contained in a network of tissues known as laticifers, located on the *Hevea* bark. Figure 1.1.2 represents a cross-section of *Hevea* bark and outlines the location of laticifer (next to the cambial tissues). As the

turgor pressure of laticifer cell is 10-15 atmospheres, when the tissue is ruptured, latex exudes from the wounding point (Hao and Wu, 2000, D'Auzac and Jacob, 1989). While some of the latex flows out, the nucleus, the endoplasmic reticulum and the mitochondria are retained close to the cell wall (Southorn and Yip, 1968). It is the retention of these key cellular components that enables regeneration of the latex, including its rubber content, in the laticifers even after a significant amount of latex is lost through tapping (Cockbain and Southorn, 1962). Therefore, the anatomical features of the laticifer cells facilitates non-destructive harvesting of latex.

Latex is harvested by the tapping process, where the outer tissues of *Hevea* bark and about 1 mm of the cambial tissues are removed (Abraham and Hashim, 1983). Following the tapping process, latex flows into an open cup, and the flow normally stops after three to five hours (Abraham and Hashim, 1983). Latex flow ceases when the wounded sites of the laticifer are sealed by the formation of rubber plugs (Abraham et al., 1975, Yeang, 1986, Sookmark et al., 2002). Despite the frequent wounding of the bark, *Hevea* can maintain its capacity to regenerate rubber consistently for 25 – 30 years (Malaysian Rubber Board, 2009, Priyadarshan, 2017a). This is because minimal removal of bark tissues ensures continuous regeneration of new bark tissues by cambial differentiation to replace the portion that was removed during the tapping process (Pakianathan et al., 1982).

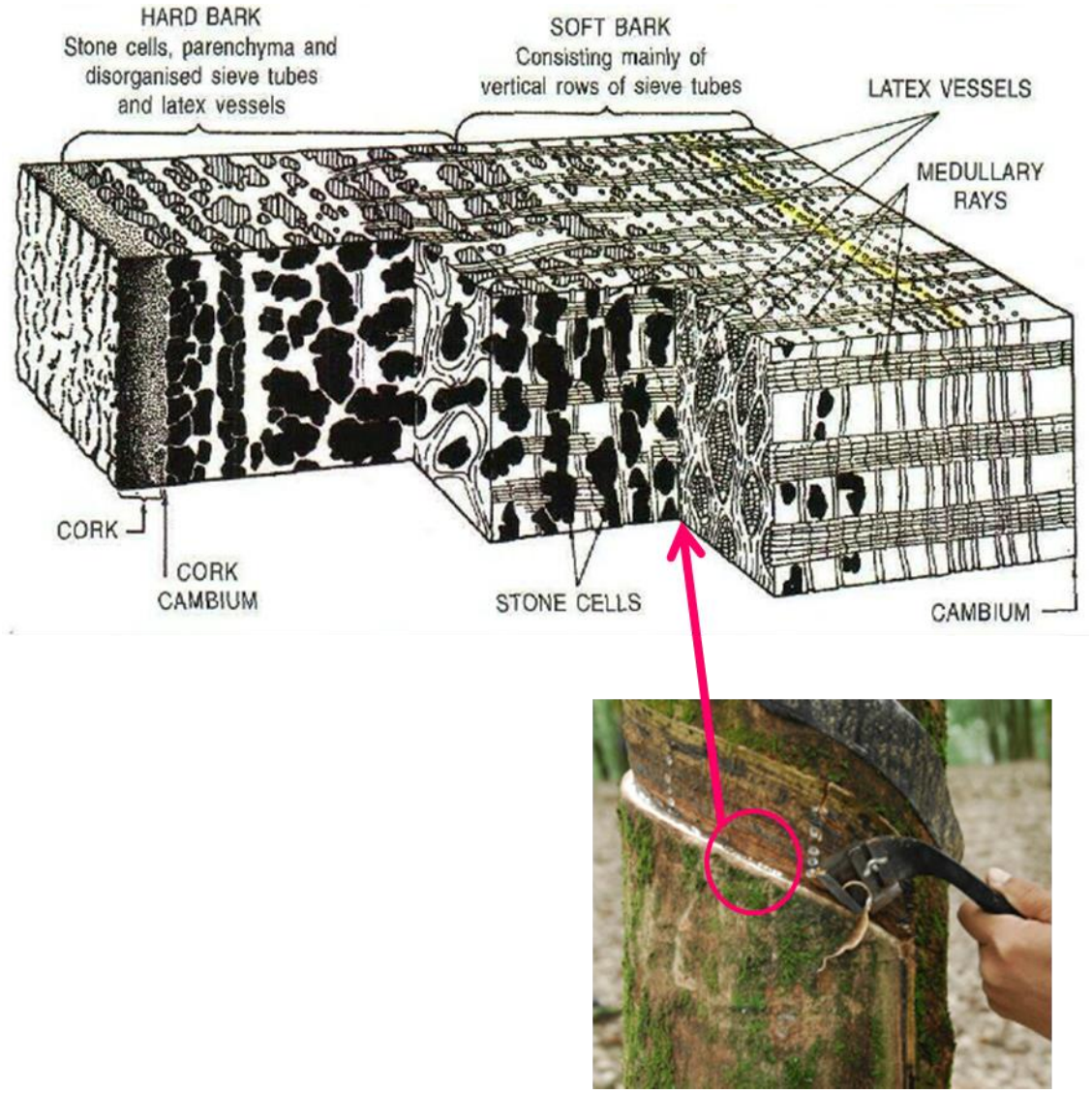


Figure 1.1.2: A schematic three-dimensional diagram of bark structure of *Hevea brasiliensis.* Harvesting of latex commences by shaving of the hard bark (about 1 mm to 5 mm thick) to wound the latex vessels (or laticifer tissues). The wounded laticifers will cause latex to flow out of the rubber tree bark. The cross section figure is depicted from Riches and Gooding (1952).

1.2. Composition of *Hevea* latex

Latex is a milky sap produced in the laticifer tissues, within *Hevea* bark. The unique feature of this exudate is that it serves as the cytoplasm of laticifer cells and hence, contains the typical cytoplasmic content of plant cells (Dickenson, 1969, Singh et al., 2003). There are three main particles suspended in latex, rubber particles, lutoids and Frey-Wyssling (FW) particles (Archer et al., 1969, Gomez and Samsidar, 1989, Gomez, 1990).

Moir (1959) was the first to demonstrate that these particles can be separated into three prominent layers when using high-speed centrifugation (59,000 g) (Figure 1.2.1). The top layer (also known as rubber phase or rubber fraction) of the centrifuged latex is formed by rubber particles. The aqueous phase of latex forms the transparent middle layer (also known as C-serum), and the bottom fraction comprises of FW particles and lutoids. In some cases, degraded FW particles accumulate below the rubber fraction.

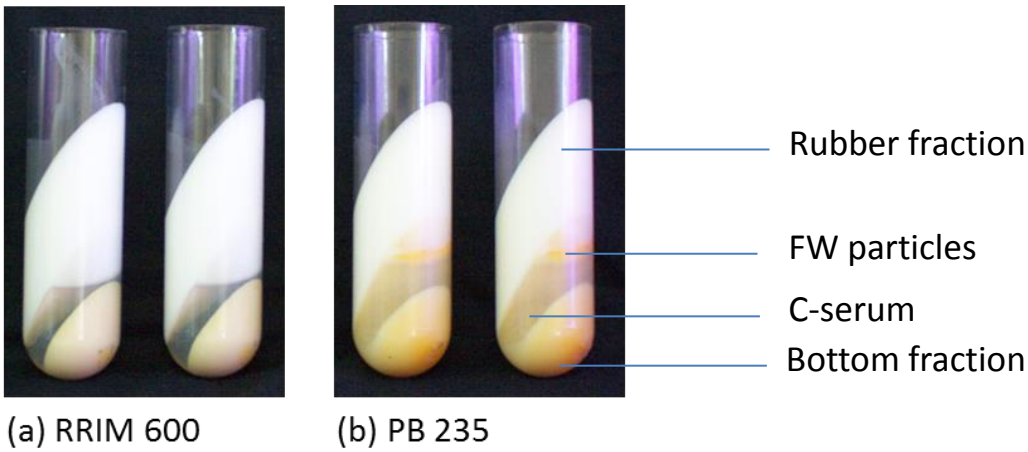


Figure 1.2.1: Latex samples collected from *Hevea* genotypes with contrasting latex colours, (a) RRIM600 (white latex) and (b) PB235 (yellow latex). For a more apparent contrast, the latex samples were centrifuged (59,000 g, one hour). The latex was separated into three different phases; the top layer known as rubber fraction, the middle layer known as C-serum and the supernatant known as the bottom fraction. Intact FW particles will form a yellow thin layer beneath the rubber fraction. In addition, another yellow layer is visible in the bottom fraction, due to the sediment of FW particles (Gomez and Samsidar, 1989).

1.2.1. Rubber particle

The rubber particle is an organelle containing a rubber core, encapsulated by a monolayer of protein and lipid membrane (Cockbain and Philpott, 1963, Cornish et al., 1999, Nawamawat et al., 2011). Rubber particles account for 30-50% of the latex total volume and hence are the most abundant latex particles (Jacob et al., 1993). There are two types of rubber particle, large rubber particles and small rubber particles. The large rubber particles commonly show a diameter larger than 0.40 μm while the small rubber particles have a diameter on average less than 0.35 μm (Ohya et al., 2000, Sando et al., 2009).

The study of rubber particle physiology and function has long been a subject of immense interest as rubber particles have been reported to be the main factor in DRC value determination (Cockbain, 1953, Chen, 1981, Cornish et al., 1999, Yamashita et al., 2018). Rubber particles are spherical particles containing a hydrophobic rubber core, encapsulated by a monolayer of phospholipid membrane in which many proteins are embedded (Dai et al., 2013, Wang et al., 2018). Important proteins associated with the rubber particle surface include rubber elongation factor (REF), small rubber particle protein (SRPP), *cis*-prenyltransferase (CPT) and rubber stimulator protein.

It has been demonstrated that small and large rubber particles have different biochemical characteristics. Yeang et al. (1995) and Sakdapipanich et al. (1999) have shown that the core of small rubber particles contains higher molecular weight rubber than that of the large rubber particles. Also, the activity of CPT is higher in the small rubber particles compared to that of the large rubber particles (Yamashita et al., 2018). Yeang et al (1995) questioned whether the biogenesis of rubber particles starts with small rubber particles

which enlarge into large rubber particles. The answer remains elusive as no experimental evidence has been reported to lend it credence.

However, recent work has drawn an analogy between rubber particle structure and that of lipid droplets and it has been reviewed extensively by (Berthelot et al., 2014a, Laibach et al., 2018). Similar to rubber particles, lipid droplets are surrounded by a phospholipid monolayer into which many proteins are adsorbed, with a hydrophobic core consisting of triacylglycerols and steryl esters (Murphy and Vance, 1999). In fact, a lipid membrane-associated protein found in avocado (lipid droplet protein, LDP) was found to have high similarity to REF and SRPP sequences (Horn et al., 2013). As such, Yamashita et al. (2016) and (Brown et al., 2017a) have explored the correlation of rubber particle biogenesis to that of lipid droplet formation. Lipid droplets were inferred to grow out toward the cytoplasm as buds and may subsequently, be separated from the endoplasmic reticulum (Chapman et al., 2012). However, the mechanism of rubber particle biogenesis and how rubber is formed on the rubber particle surface remains unclear.

1.2.2. Frey-Wyssling particle

The FW particles were first described by Frey-Wyssling (1929), and they exist as spherical, double-membrane globules (Dickenson, 1969, Southorn, 1960, Gomez and Samsidar, 1989). Their distinctive yellowness is attributed to the presence of carotenoids, causing the yellow tinge in some latex samples (Gomez and Moir, 1979). FW particles form 15% to 35% of the latex total volume (D'Auzac and Jacob, 1989). The presence of carotenoids indicates that FW particles may contain enzymes that catalyse the formation of isoprenoid

products (Wititsuwaannakul et al., 2004). However, little information is available to ascertain the suggestion of FW particles' role in the isoprenoid biosynthesis pathway.

1.2.3. Lutoid

The lutoids are spherical membrane-bound vacuoles, containing acidic hydrolases (Zakia et al., 1992). Within lutoids, a wide range of non-rubber substances such as proteins (e.g. hevein), enzymes (for example chitinase and β-glucanase) and metabolites have been identified (Jacob et al., 1993, Wititsuwannakul et al., 2008, Wang et al., 2013, Subroto et al., 1996). Lutoids have been hypothesised to contribute to the viscosity and colloidal properties of latex and hence have been inferred to play a role in latex flow. This is due to the released of cations, proteins and enzymes from the lutoids which is associated with latex destabilisation (Subroto et al., 1996, Gomez, 1990, Wititsuwannakul et al., 2008). Substantial evidence indicates that during the tapping process, lutoids are ruptured and release their contents into the latex (Yeang, 1986, Sunderasan et al., 2007). The released contents destabilise the surrounding rubber particles, lead to the formation of rubber flocculation and hence latex coagulation (Wititsuwannakul et al., 2008).

1.2.4. Non-rubber constituents

Although variation occurs in the composition of latex from different trees due to genetic and environmental factors, the typical percentages of the major latex constituents have been are summarised in Table 1.2.4.1. Of the several

non-rubber compounds, most investigation has focused on the proteins and lipids. This is due to their apparent roles in rubber formation and their effect on processed rubber quality.

Table 1.2.4.1: Typical composition of fresh latex collected from *Hevea brasiliensis.* The latex composition was obtained from work reported by D'Aujac and Jacob (1989).

Component	**Percentage (%)**
Rubber	25-50%
Protein	1-1,8
Carbohydrates	1-2
Neutral lipids	0.9-1.6
Inorganic constitutents	0.6-0.6
Amino acids, amines	0.4
Water	about 59

Archer and Audley (1987) demonstrated that 66% of the total latex protein is located in the latex serum, 20% localised on the rubber surface and the remaining 14% is found on lutoids or FW particles (collectively known as bottom fraction). Proteomic profiling by Wang et al. (2018) further supports previous findings, where 1,837 proteins were reported in latex serum, 1,739 and 1,020 proteins were found in rubber particle and bottom fraction, respectively. As mentioned previously, latex proteins are implicated in rubber formation and impact on rubber quality. For example, rubber particle proteins such as REF, SRPP, CPT and rubber stimulator protein have been demonstrated to enhance the uptake of IPP into the rubber molecule *in vitro* (Light and Dennis, 1989, Yusof et al., 2000, Oh et al., 1999, Asawatreratanakul et al., 2003, Yamashita et al., 2018). In addition, Hasma and Alias (1990) demonstrated that the retention strength of processed rubber is inversely correlated to the levels of accumulated protein in latex.

Ho et al. (1975) investigated the fractionation of latex total lipids using thin-layer chromatography. They found the total lipids in latex can be divided into

three groups, namely neutral lipids, glycolipids and phospholipids. Latex total lipids are insoluble in water and hence, the components are associated with the surface of the particles inside latex. Later, Hasma and Subramaniam (1986) used a gas chromatography approach to characterise the total lipids in the latex of *Hevea brasiliensis*. They identified neutral lipids to consist of carotenoids, sterol, triglycerides, free fatty acids, free fatty alcohols and diglycerides while free and esterified tocotrienols and fatty alcohol acetates make up the glycolipid content and finally, the phospholipid fraction contains mainly phosphatidyl ethanolamine, phosphatidyl choline and phosphatidyl inositol. Amongst these total lipids, neutral lipid was found to have a higher impact on latex stability. Previous findings indicated that higher amounts of neutral lipids increased the likelihood of coagulating latex samples (Sherief and Sethuraj, 1978, Liengprayoon et al., 2013). Total carotenoids from the neutral lipids have been identified as the primary pigments that result in a darker processed latex colour (Verhaar, 1959, Sakdapipanich, 2006), which can be regarded as a negative characteristic.

The carbohydrate content in latex does not appear to influence the technical properties of processed rubber. Bealing (1969a), (1969b) found that glucose, fructose and sucrose are the predominant free sugars in *Hevea* latex. It is presumed that these sugars are involved in the formation of precursors for rubber synthesis. The expression level of the sucrose synthase gene (encoding an enzyme that involved in sucrose metabolism) shows a positive correlation with latex yield in *Hevea* (Dusotoit-Coucaud et al., 2009, Xiao et al., 2014). However, experimental evidence for the incorporation of sugars into rubber in *Hevea* latex when it was incubated with ^{14}C-labelled sucrose, fructose or glucose (D'Auzac and Jacob, 1989) is lacking.

The mineral content (or inorganic constituents) of latex was inferred to play roles in rubber elongation steps (Cornish, 2001) and latex stability (Morris and Lakin, 1995). For example, magnesium ions are a prerequisite co-factor for the uptake of IPP into rubber molecules (Archer et al., 1963, Cornish et al., 1999). On the other hand, a high ratio of magnesium to phosphate ions is often found in latex with a higher propensity to coagulate (Hasma and Alias, 1990).

1.3. Rubber biosynthesis

Rubber in *Hevea* latex is defined as an isoprenoid as it possesses a long sequence of isoprene repeats, as shown in Figure 1.3.1 (Tanaka, 1985, Sakdapipanich, 2007). Thus, rubber shares a precursor known as isopentenyl diphosphate (IPP) with other non-rubber isoprenoids in latex. It has been established that there are two biosynthetic routes involved in providing IPP in higher plants; the cytosolic mevalonate (MVA) pathway and the plastidic methylerythritol phosphate (MEP) (Lichtenthaler, 1999). The established routes of the MVA and the MEP pathways in generating IPP are shown in Figure 1.3.2.

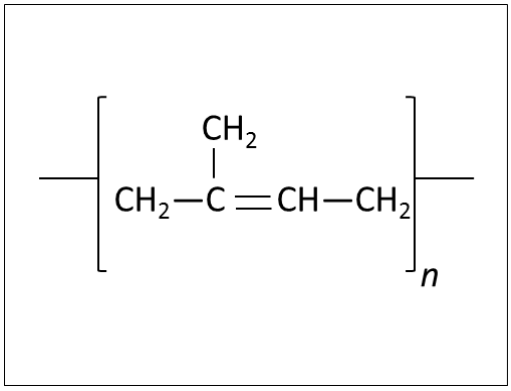


Figure 1.3.1: Chemical structure of natural rubber main monomer, known as isoprene or 2-metyl-1,3-butadiene. The isoprene is constructed from isopentenyl diphosphate (IPP) and its isomer, dimethylallyl diphosphate (DMAPP). Long chain of natural rubber (consisted of many repeats, *n* of the monomer) is located inside rubber particle, suspended in the latex of *Hevea brasiliensis.*

The MVA route starts with the condensation of acetyl coenzyme A (acetyl Co-A) molecules and ends with the formation of IPP and its isomer, dimethyl allyl diphosphate (DMAPP). Early work in the biochemistry of rubber biosynthesis by Archer et al. (1963) demonstrated that acetate in the MVA pathway was incorporated into rubber, thus supporting the MVA route for IPP for rubber formation. Hepper and Audley (1969) were the first researchers to demonstrate hydroxy-3-methylglutaryl coenzyme A (HMG coA), which is a precursor from the MVA pathway, and acts as the rate-limiting step for IPP incorporation into the rubber chain. Later, this observation was corroborated by the expression analysis of transcripts encoding hydroxy-3-methylglutaryl-coenzyme A reductase (HMGR). HMGR is an enzyme catalysing the production of HMG coA in the MVA pathway. The *HMGR* transcript level was found to be elevated in rubber-producing plants that displayed higher rubber content (Sando et al., 2008a, Pütter et al., 2017).

Although the MVA pathway has conventionally thought to be the provider of IPP. recent work on the MEP pathway suggests that this route may serve as an alternative provider of IPP for rubber formation (Sando et al., 2008b, Chow et al., 2012). The MEP pathway starts with the condensation of pyruvate and D-glyceraldehyde-3-phosphate and is succeeded by six sequential steps resulting in IPP generation. However, conclusive proof demonstrating the incorporation of the MEP-generated intermediate into rubber molecule is lacking.

Following the formation of IPP, there are three steps that are unique to rubber formation, namely initiation, elongation and termination. The initiation step involves the incorporation of IPP and isoprenoid intermediates (geranyl diphosphate; GPP, farnesyl diphosphate; FPP, and geranylgeranyl

diphosphate; GGPP), mediated by rubber transferase to create short hydrophobic chains (Archer and Audley, 1987, Cornish and Siler, 1995). Once rubber biosynthesis has been initiated, the hydrophobic chain is sequestered inside the rubber particle (Puskas et al., 2006). The chain elongation is assumed to occur with the subsequent incorporation of IPP into the existing rubber molecules inside rubber particles (Cornish et al., 1999, Chiang et al., 2014). Based on the investigation of rubber structure and rubber formation *in vitro*, it was found that the possible isoprenoid intermediate species used in the initiation step are either FPP or geranyl diphosphate GPP(Archer and Audley, 1987, Tanaka et al., 1996). However, Cornish (2001) demonstrated that FPP is the predominant isoprenoid intermediate for the initiation process. As it has been established that the MVA route produces FPP while the MEP pathway provides GPP and GGPP (Chappell, 1995, Rohmer, 1999, Hemmerlin et al., 2012), the findings reported by Archer et al (1987) and Tanaka et al (1996) have supplied the first experimental evidence of possible cross-talk of the cytosolic IPP and the plastidic IPP. Indeed, recent work using a transcriptome-based approach inferred that there might be flux of the plastidic IPP into the latex cytosol for rubber formation (Chow et al, 2012).

In *Hevea brasiliensis* latex, isoprenoid products can be divided into rubber and non-rubber isoprenoids. The non-rubber isoprenoid such as carotenoids (Hasma and Subramaniam, 1986, D'Auzac and Jacob, 1989), tocotrienols (Dunphy et al., 1965), tocopherols (Whittle et al., 1967), plastoquinone (Phatthiya et al., 2007) and dolichols (Tateyama et al., 1999) are relatively smaller in amount compared to rubber in latex. Despite the smaller proportion, these compounds require IPP for their formation. However, detailed information concerning the contribution of cytosolic IPP or plastidic IPP towards isoprenoid

synthesis in latex is yet to be reported. A growing body of evidence indicate that the MVA pathway was shown to be the sole provider of rubber formation in *Hevea* young seedlings (Sando et al., 2008a). However, no data being reported for the IPP utilisation in the mature rubber trees.

Previous breeding data documented that the productivity of latex is positively correlated to *Hevea* tree age, indicating a lower rubber biosynthetic rate occurs in *Hevea* seedlings (Ong et al., 1995, Priyadarshan, 2017b). The variation of rubber biosynthesis rate between the seedlings and mature *Hevea* trees led Chow et al (2012) to propose that the MEP pathway may contribute to the IPP pool for rubber formation in mature *Hevea* trees, to meet the demand of precursor needed by the rubber biosynthetic route. Nevertheless, the finding was inferred based on the expression analysis of genes involved in the MVA and MEP pathway and therefore needs further experimental evidence to support the hypothesis.

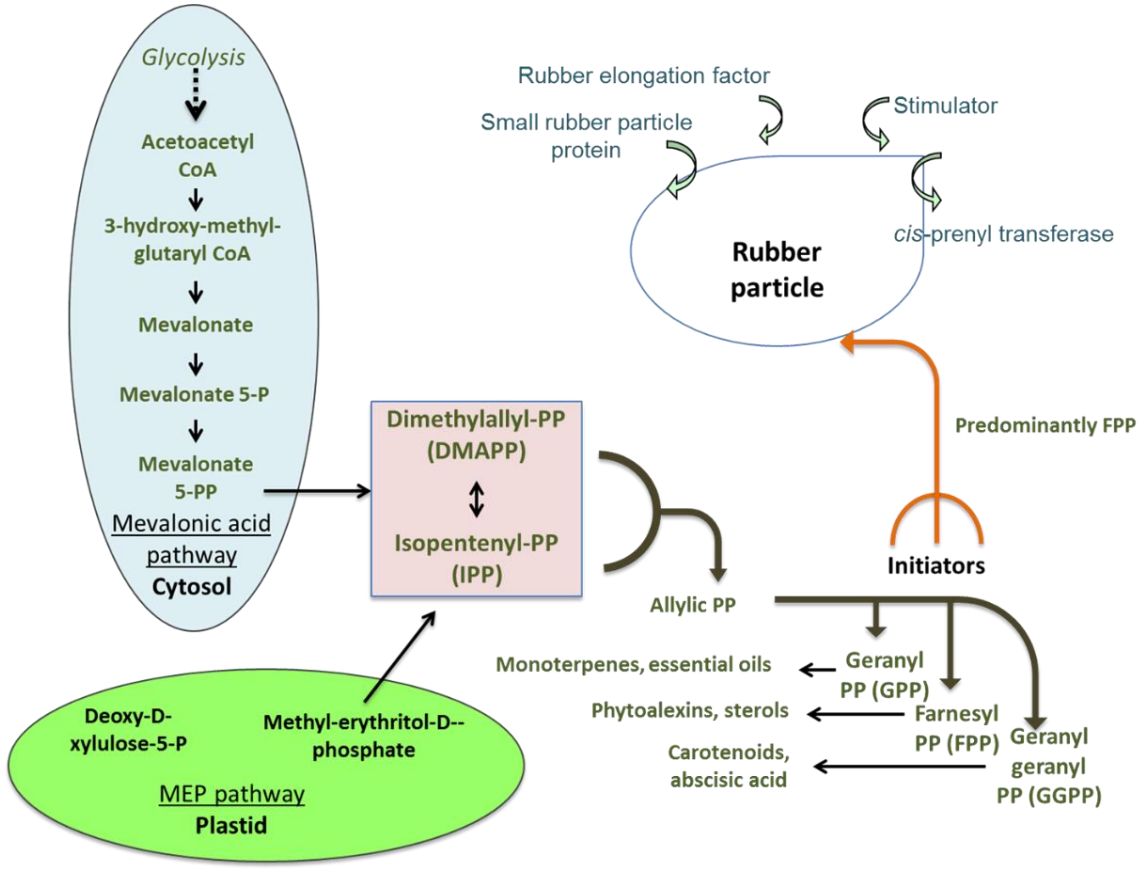


Figure 1.3.2: A schematic diagram modified from Cornish and Blakeslee (2011), showing the isoprenoid biosynthesis pathways in plant cells. The same scheme is also most likely to occur in the latex of *Hevea brasiliensis.*

A model describing steps for rubber elongation in rubber-producing plants has been proposed by Backhaus and Bess (1985). The model involves rubber transferases (that are bound to the surface of the rubber particles) using IPP into the growing rubber molecules inside the rubber particles. CPT was observed to exhibit rubber transferase activity *in vitro* when it formed a protein complex with a protein called *Hevea* rubber transferase-rubber elongation bridging protein (HRBP) (Yamashita et al, 2016). Brown et al (2017) have further shown that REF interacts with SRPP, to mediate the localisation of the CPT-HRBP protein complex onto the rubber particle surface. The interaction of the protein complex with REF and SRPP is presumed to be involved in the rubber elongation steps as proposed by Backhaus and Bess (1985). The uptake

of IPP into the pre-existing rubber molecule leads to the elongation of the high molecular weight cis-polyisoprene molecules (Steinbüchel, 2003).

The termination of rubber formation is presumed to occur when the rubber transferase complex dissociates from the growing rubber molecule (Amerik et al., 2018). Previous findings revealed that rubber transferase activity is inversely proportional to the total surface area of a rubber particle (Archer et al., 1963, McMullen and McSweeney, 1966). It is then suggested that as the rubber particle size increases in volume, the enzyme distribution on the surface is reduced and thus, the specific activity per unit surface area is decreased (Yusof and Chow, 2003). This leads to the impedance of rubber formation. However, direct evidence to describe the mechanisms that underlie the initiation and termination of rubber biosynthesis is still lacking.

1.4.1. *Hevea* genotype improvement for desired traits

Latex is the main product of *Hevea* cultivation and thus crop improvement efforts are focused towards increasing the production of latex. Apart from latex yield, emphasis is also given to the secondary characteristics that include bark volume, resistance to wind damage, leaf and bark diseases and tapping panel dryness (Ong et al., 1994, Priyadarshan, 2017b). In Malaysia, the main *Hevea* breeding programme is carried out by the Malaysian Rubber Board. It began in 1928 and continues currently. During this time, the rubber breeding programme has undergone seven phases of crop improvement. This has resulted in seven series of RRIM clones, ranging from RRIM 500 to RRIM 3000 (Table 1.4.1.1) (Ong et al., 1994, Nurmi-Rohayu, 2017). The breeding of *Hevea* has elevated

the latex yield from around 400 kilograms of rubber per hectare per year (kg/ha/yr), to the current experimental yields of about 3000 kg/ha/yr (Burkill, 1959, Ramli et al., 2005).

Table 1.4.1.1: Timeline of *Hevea* breeding programme carried out by the Malaysian Rubber Board.

Phase	**Rubber tree clone series**	**Period of breeding**	**High yielding clone**
Pre-breeding period	Seedling	Early 1920s	No info
I	RRIM 500 series	1928 – 1931	RRIM 501
II	RRIM 600 series	1937 – 1941	RRIM 600, RRIM 605 and RRIM 623
III	RRIM 700 series	1947 – 1958	RRIM 701 and RRIM 712
IV	RRIM 800 series	1959 – 1965	RRIM 803
V	RRIM 900 series	1966 – 1973	RRIM 901, RRIM 928, RRIM 937
VI	RRIM 2000 series	1974 – 1980	No info
VII	RRIM 3000 series	1981 – now	RRIM 3001

Initially, crop improvement was performed through planting of seeds produced from superior *Hevea* seedlings. The examples of *Hevea* clones produced from this method include Pil D 84, Pil D 65 and PB 86 (Mann, 1934). Afterwards, starting from 1928, a conventional breeding programme involving controlled cross-pollination of superior parents was begun. Conventional *Hevea* breeding involves crossing of current elite cultivars with the donors of useful properties, followed by the selection of superior recombinants (Figure 1.4.1.1).

In a general plant breeding programme, the process includes several crosses and several generations, requiring careful phenotypic selection. However, in *Hevea*, the crossing process only involves the generation of the first filial (F1) hybrids, and the poor-performance siblings are culled during preliminary screening. The remaining seedlings that make the cut are then propagated through bud-grafting (involving scion of the evaluated progeny grafted onto rootstock prepared from *Hevea* seedlings) and subjected to clone assessment. The clone appraisals include two consecutive trials, starting with a small-scale clone trials, followed by a large-scale clone trials of potentially valuable lines. The most commonly used planting design for the clone trial is either randomised block or a balanced simple lattice, with three to five replicates (Ong et al., 1994). Finally, outstanding clones from the clone trials are recommended for industry uptake.

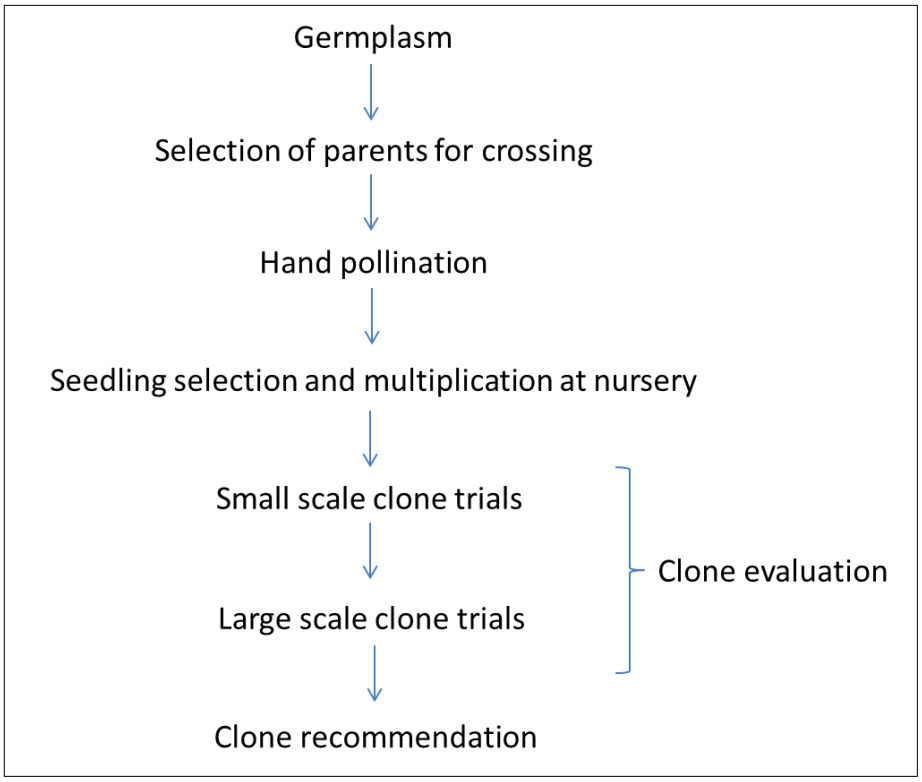


Figure 1.4.1.1: A typical breeding cycle for rubber crop improvement. In the small-scale clone trial, the clones (propagated seedlings) are planted in a block planting (usually lattice design) with 8 trees per plot. Each plot is replicated twice. The evaluations include rubber yield, branching habit, disease resistance, bark thickness and bark renewal. Usually, the time taken for a complete cycle of a small-scale clone trial is approximately 11 years. Promising clones are planted in larger plots for large scale clone trials. The large-scale trial usually takes 16 to 18 years to complete (Malaysian Rubber Board, 2009).

1.4.2. Issues in *Hevea* breeding

Being a perennial tree crop, breeding of *Hevea* requires considerable time and high operating cost. Indeed, a complete cycle of *Hevea* breeding takes about 30 years (Tan, 1978, Simmonds, 1985, Priyadarshan, 2017b). In addition, *Hevea* needs between 4 to 5 years to reach maturity and show a steady stream of latex production. The appraisal of new progenies usually involves a large experimental area and considerable labour resources. For example, 20,000 to 30,000 hand pollinations are performed yearly in the Malaysian Rubber Board breeding program, and only few clones would be selected at the end of a breeding cycle (Malaysian Rubber Board, 2009).

There have been considerable efforts to shorten the length of *Hevea* breeding, particularly during the selection cycle. Recent advance in genetic marker technology offers tools for rapid characterisation of genotypes based on their genetic content. The polymorphic markers such as isozymes, random amplified polymorphic DNA (RAPD), restricted fragment length polymorphism (RFLP), microsatellite loci and single nucleotide polymorphism (SNP) have been developed for *Hevea* (Yeang and Chevallier, 1999, Besse et al., 1994, Lespinasse et al., 2000, Le Guen et al., 2009, Pootakham et al., 2015, Conson et al., 2018). These markers have been used to explore the genetic diversity of *Hevea* populations, genotype identification and generation of quantitative trait loci (QTL) (Besse et al., 1994, Rao et al., 2013, Safiah, 2012). These genetic markers can be utilised in the early screening of potential progenies at the nursery level. This will maximise the number of potential progenies in the selection cycle and therefore lead to a possible increase in the number of new rubber tree genotypes recommended for planters.

1.5 Rubber genomic and transcriptomic studies

The molecular events of IPP formation from the MVA and the MEP pathway were well documented before the advent of Next Generation Sequencing Technologies (NGS). Sando et al (2008a, 2008b) characterised all biosynthetic genes implicated in both routes through cDNA cloning and functional characterisation. In addition, proteins and genes associated in the final steps of rubber elongation namely REF, SRPP, CPT and rubber stimulator (Kush et al., 1990, Chow et al., 2007, Chow et al., 2006, Asawatreratanakul et al., 2003, Attanyaka et al., 1991) were also characterised. Analysis of proteins and transcripts were exclusively used at that time to identify genes of interest., Expressed sequence tags (EST) for *Hevea* latex were constructed by Han et al. (2000), Ko et al. (2003), Chow et al. (2007). Their findings revealed specific genes encoding proteins associated with rubber particles and involved in rubber formation.

Recently, increased resolution of the molecular events underlying rubber biosynthesis in *Hevea* latex had been achieved through NGS approach (Mantello et al., 2014). The utilisation of transcriptomic profiling has shown that most of the key genes involved in rubber biosynthesis are encoded by multigene families. Furthermore, some of these key genes showed sub-functionalisation in different tissues (Tang et al., 2016, Lau et al., 2016). Although the *Hevea* genome is highly heterozygous and contains a significant proportion of repetitive regions, the first public draft version was reported by the Malaysian Rubber Board in 2014 (Mollison et al., 2014). Afterwards, three other draft genome versions were reported based on the genotypes RRIM600, BPM24 and Reyan7-22-97 (Tang et al., 2016, Lau et al., 2016, Pootakham et

al., 2017). Analysis of genome assemblies revealed that *Hevea* has significant homology with cassava in the same family. This is due to the fact that both *Hevea* and cassava shares an ancestral whole genome duplication (Bredeson et al., 2016, Pootakham et al., 2017). Indeed, based on the estimation of the divergence of *Hevea* and cassava using a fossil-calibrated molecular clock approach, it is predicted that the common ancestral diverged about 35 to 36 million year ago (Bredeson et al, 2016, Pootakham et al, 2017).

1.6 Rubber metabolite studies

The principal focus of metabolite-based studies in *Hevea* in the past revolved around non-rubber constituents. As described previously, total lipids in latex have been characterised by Ho et al (1975), Hasma and Subramaniam (1986), Sakdapipanich et al (2006). The isolation and characterisation of these compounds using chromatography techniques were pursued because though present in small amounts, these compounds have a pronounced effect on the properties of rubber.

Hevea latex colour is occasionally observed to be yellow due to carotenoid content, which accumulates in FW particles. Eaton and Fullerton (1929) were the first to identify carotenoid as the primary pigment in the processed latex. This indicates that carotenoids are retained throughout the rubber processing step. More recently, the main carotenoid pigment in fresh latex and processed rubber was identified as β-carotene by Sakdapipanich (2006) and Liengprayoon et al. (2013). The presence of carotenoids is regarded as a non-desirable characteristic as it causes darkening of processed rubber.

Different isoprenoid initiators were reported to be involved in rubber and carotenoid synthesis. Based on the incorporation of ^{14}C-labelled isoprenoid diphosphates into the rubber molecules, it was shown that FPP is the preferred initiator for rubber synthesis (Cornish, 2001). On the other hand, the first committed step of carotenoid biosynthesis is the condensation of geranylgeranyl diphosphate (GGPP) molecules to form phytoene, mediated by phytoene synthase (PSY) (Cazzonelli and Pogson, 2010). All allylic isoprenoid diphosphates (GPP, FPP and GGPP) are built from IPP. Currently, the mechanism that determines the IPP flow to rubber and carotenoid synthesis has not been ascertained. In this regard, information on the various competitive IPP sinks in latex might help in developing strategies to enhance IPP supply directed towards rubber biosynthesis while reducing competition from non-rubber isoprenoids.

1.7 Challenges in *Hevea* biological studies

Hevea brasiliensis originated from the Amazonian rainforest and hence, it exhibits typical tropical tree characteristics. For example, mature *Hevea* trees can reach a height of 20-30 metres and commonly planted in a large are of plantation field. However, it is only under such field conditions that realistic assessment of factors influencing yield and quality of *Hevea* rubber can be obtained. In field situations, *Hevea* trees might vary due to the environmental or microclimatic conditions. As such, field materials may show high levels of heterogeneity which can influence the results of any transcriptome or metabolome profiling. In addition, *Hevea* trials occupy a large planting acreage, and this posing logistical challenges in transporting the collected

samples from planting sites. Even with an organised field management and supported by experienced handlers, mis-identification of planting materials is not uncommon in forest tree cultivation (Harju and Muona, 1989, Wheeler and Jech, 1992, Kawauchi and Goto, 1999, Asif et al., 2017) and there are logistic difficulties in transferring biological samples from the field to the laboratory. As the clone trials span decades, differences in genotype composition of trials and planting design among large scale clone assessments was not surprising. For example, in the Malaysian Rubber Board breeding programme, most of the clone trials for phase I and phase II were based on the performance of *Hevea* genotype RRIM501 as control (Burkill, 1959), phase III-IV used RRIM600 and phase V onwards employed PB 260 and RRIM 600 (Table 1.4.1.1). In some cases, planting of old recommended genotypes has been discontinued due to diseases. Such discrepancies have created anecdotal data recording, which subsequently make comparison between trials challenging.

Latex is a complex and relatively unstable material to handle and transfer from the field to stable storage in the laboratory. During the tapping process, some of the lutoids are destroyed and the content released from the ruptured lutoids leads to the onset of rubber flocculation (Yeang, 1986). Also, sugars within the harvested latex are subject to bacterial degradation, which will create a shift in pH of the latex (Woo, 1973). These two conditions cause harvested latex to gradually thicken and finally coagulate. Therefore, any investigation involving fresh latex is challenging because of its coagulating tendency. Appropriate collecting strategies must be considered so that the metabolic state of the samples at the time of its collection can be preserved.

1.8. Research aims and approach

Since the inception of rubber breeding programmes, improving rubber content in the latex of *Hevea brasiliensis* has always been one of the main aims. As such, an understanding of rubber biosynthesis and the relationship of latex composition to rubber quality are topics of immense interest in *Hevea* breeding. A complete understanding of biosynthetic pathway that provides IPP for rubber and non-rubber isoprenoids might help in the future genetic manipulation of *Hevea brasiliensis*.

Past research has contributed substantially towards understanding how the *Hevea* laticifers make rubber (Lynen, 1969, Bandurski and Teas, 1957, Archer and Audley, 1987, McMullen and McSweeney, 1966). Work on latex physiology has revealed that some non-rubber components are retained in the processed latex and hence affect the physical properties of the final product (Archer and McMullen, 1960, Subramaniam, 1980).

However, there are gaps in genetic information that regulates rubber biosynthesis. Although key genes involved in the biosynthetic steps leading to rubber and non-rubber isoprenoid synthesis have been reported, for many definitive gene models or their final genome locations are yet to be confirmed. The lack of a pseudomolecule quality genome sequence currently impedes the construction of such resources. The assembly of a complete *Hevea* genome is highly challenging as the short-read sequences used to construct the genomic scaffolds could not resolve the repetitive genomic regions. It has been estimated that 68-70% of rubber genome is of repetitive DNA regions (Low and Bonner, 1985, Lau et al., 2016, Tang et al., 2016, Pootakham et al., 2017, Mollison et al., 2014, Leitch et al., 1998). It is only very recently that the

availability of new molecular and cytological tools to generate a high-quality rubber genome becoming available. Furthermore, although darkened rubber does not negatively impact its physical properties, processed rubber is classified on the basis of colour. Premium processed rubber on the market is very light and uniform in colour and any departure from this characteristic is detrimental to the rubber price. Therefore, from a commercial point of view, an understanding of the control of latex carotenoids would be of significant value. .

In the present study, the transcript levels of the genes involved in the isoprenoid biosynthesis of both rubber and carotenoid were investigated in the latex of two *Hevea* genotypes (RRIM600 and PB235) that have visibly contrasting carotenoid constituents. The project has four objectives, outlined as follows:

1. To analyse the levels of carotenoid in *Hevea brasiliensis* latex of RRIM600 and PB235.
2. To develop methods to investigate the content of isoprenoid intermediates in the latex of RRIM600 and PB235 *Hevea brasiliensis*.
3. To construct a reference transcriptome to underpin *Hevea* transcript profiling.
4. To demonstrate the utility of the constructed reference transcriptome through the expression profiling of RRIM600 and PB235 latex samples and the characterisation of a key gene involved in rubber formation.

1.9. Research summary

The subjects investigated are related and thus, the thesis has been organised as follows to provide overall coherence. The thesis is divided into two parts, metabolite characterisation and transcriptome profiling. In the metabolite characterisation, analytical method development for the identification and quantification of the targeted isoprenoids from plant extracts are described. Transcriptome profiling part concerns firstly the construction of reference transcriptome and secondly, the utilisation of genome and transcriptome resources in measuring key genes of the isoprenoid biosynthetic pathway. Consequently, a pilot characterisation study on the genomic region bearing multiple copies of gene encoding REF and SRPP (rubber particle-associated proteins) was carried out.

Chapter 2

Materials and Methods

2.0. Background of the plant materials and sampling strategy

The main set of *Hevea* trees used in this research were planted in year 1993, in Field 23, in a simple planting block design (8x5 rows) at the Rubber Research Institute of Malaysia Experiment Station, Malaysia. The plot was maintained according to the standard agronomic practices for rubber plantations, as recommended by the Malaysian Rubber Board (Malaysian Rubber Board, 2009). At the Malaysian Rubber Board, it is a practice to ensure that all *Hevea* trees are subjected to a systematic tapping known as '1/2S d/2' tapping system. In this tapping system, the trees are tapped with a half spiral downward cut, every two days (Figure 2.0.1 (b)) (Malaysian Rubber Board, 2009). The tapping process is stopped on rainy days.

Hevea tree have a high turnaround of latex metabolism. They can quickly regenerate latex after the tapping process removes some of the cytoplasmic content (Gomez and Moir, 1979, Jacob et al, 1993). In addition, as long as the tapping process does not remove too much of the cambial tissues, the renewal of laticifer tissue continues unhindered (Abraham et al., 1975, Pakianathan et al., 1982). A skilled tapper will tap to the correct depth of the bark without wounding the cambium (Figure 2.0.1(a)). In this study, the tapping process was performed by one of the experienced tappers from the Malaysian Rubber Board, Mr Monyrajan Venugal.

It is a common practice to add ammonia (0.05% weight/weight) to fresh latex, to preserve the stability of latex. In this study, preservation with ammonia was not applied on the collected latex so that the biological conditions of the latex at the point of collection was maintained. The collection of latex and some

of the processing steps were performed *in situ* to minimise degeneration of the samples

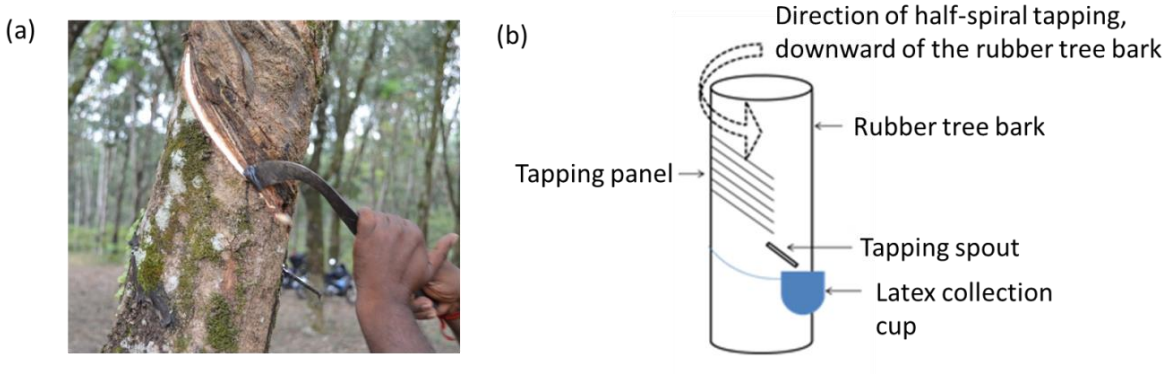


Figure 2.0.1: (a) Latex collection through incision of the *Hevea* tree bark. (b) A schematic diagram depicting the direction of the incision made on *Hevea* tree bark for latex harvesting.

2.1. Plant materials

2.1.1 *Hevea brasiliensis*

For each *Hevea* tree genotype, ten trees were randomly selected from the trial for latex collection. For each genotype, out of the ten selected trees, only six were used for the subsequent experiments (RNA extraction, metabolite extraction and dry rubber content measurement). The latex collection was performed in the morning (when the turgor pressure of the tree was highest) and the same latex samples were used for metabolite, total RNA extractions and also for the measurement of dry rubber content. Latex samples were immediately transported to the laboratory (0.8 km away from the planting plot) for the subsequent processing (as described in sections 2.4, 2.5, 2.13 and 2.14).

The *Hevea* tree genotypes RRIM600 and PB235 were selected based on the appearance of latex colour (refer to Figure 1.2.1 in Chapter 1). Latex colour

is a clonal trait (Rubber Research Institute of Malaysia, 1970, Rubber Research Institute of Malaysia, 1969), whereby some *Hevea* genotype produces white latex (such as RRIM600) and some *Hevea* tree genotype generates (namely PB235) latex with a slightly yellow tinge.

2.1.2. *Solanum tuberosum*

Solanum tuberosum cultivars Mayan Gold, Cabaret and Desiree were grown from seed tubers. *S. tuberosum* was used in the method development of hydrophilic interaction liquid chromatography (HILIC) methods. The seed tubers were obtained from Mr Ralph Wilson of the James Hutton Institute. The seed tubers were planted in pots (15 cm diameter) consisting of compost mix of 85% (v/v) Irish moss peat, 7% (v/v) Pavoir sand, 7% (v/v)Perlite, 0.2% (w/v) limestone (magnesium), 0.2% (w/v) limestone (calcium), 0.1% (w/v) Sincrostart base fertiliser (William Sinclair, UK), 0.15% (w/v) and 0.03% (w/v) Intercept insecticide (Bayer CropScience, UK), 0.15% (w/v) Osmocote mini controlled-release fertiliser (Scotts, UK), 0.1% (w/v) and Celcote wetting agent (LBS Horticulture, UK). The plants were grown in June 2016, when the average day time temperature was 20 ^{0}C and night time 15 ^{0}C. Plant tissues (leaf and tuber) were harvested when the plants reached maturity. The tissues were excised, rinsed in water, and immediately immersed in liquid nitrogen. The flash-frozen tissues were kept at -80 ^{0}C until removed for analysis.

2.2. Chemicals

All chemicals and solvents used in sample extraction, preparation and separation of carotenoids and isoprenoid intermediates were of HPLC grade, except when mentioned otherwise. Acetone, *tert*-methyl butyl ether (MTBE), chloroform, acetonitrile and ammonium acetate Optima® were purchased from Fisher Scientific (United Kingdom). Ethyl acetate HyperSolv Chromanorm® and methanol HyperSolv Chromanorm® were supplied by R&L Slaughter Ltd (United Kingdom). The standards for trans-β-Apo-8′-carotenal, β-carotene and α-carotene were purchased from Sigma-Aldrich (Switzerland). Lutein and zeaxanthin standards were purchased from Extrasynthese (France). The standards for isopentenyl diphosphate (IPP), dimethylallyl diphosphate (DMAPP), geranyl diphosphate (GPP), geranylgeranyl diphosphate (GGPP), farnesyl diphosphate (FPP), deoxy-xylulose diphosphate (DXP) and methylerythritol phosphate (MEP) were purchased from Echelon Biosciences (US). Deionised water obtained from a Purelab Option-Q water purification system (Elga, United Kingdom) was used in rinsing steps and buffer preparations.

2.3. Standards and solution preparation

For each carotenoid standard, a 1 mg/ml solution stock was prepared. A total of 1 mg of standard was dissolved by sonication in 1 ml of 80/20 (percentage; volume/volume) ethyl acetate/deionised water. For the calibration curve, stock solutions of appropriate concentrations were prepared by serial dilution of the initial stock solution. The analytical standards for isoprenoid

intermediates were prepared by adding deionised water to a final concentration of 1 mg/ml. The dissolved carotenoid and isoprenoid standards were stored at 4 ^{0}C and a working solution (100 µM of final concentration) was prepared fresh before each chromatography run.

To prepare a 50 mM ammonium acetate stock solution, 3.85 g ammonium acetate was dissolved in 1 L of deionised water. The pH of the stock solution was adjusted with 30% (v/v) ammonium hydroxide or acetic acid. Additionally, the stock was diluted using deionised water, to the final concentration of 10 mM to be used as the mobile phase buffer in the chromatogram run.

2.4. Carotenoid extraction

Optimisation of the solvent extraction was performed so that the optimal yield of carotenoid could be recovered from the latex samples. There were three parameters evaluated in the extraction optimisation: i) the selection of suitable organic solvent; ii) the determination of optimal time for the extraction incubation; and iii) the determination of extraction yield from liquid and coagulated latex.

The selection of a suitable organic solvent for carotenoid extraction was determined from a choice of acetone, petroleum ether and tetrahydrofuran. The optimisation was conducted by adding solvent into the latex samples, with the mixture ratio of latex sample to solvent at 1:5 (volume/volume). When the latex sample was mixed with solvent, the rubber particles would coagulate into a non-reversible rubber sponge-like structure. At the same time, the interaction of latex with solvent causes the membrane structure of rubber particles to partially degrade into an insoluble portion called the gel phase (Ehabe and Bonfils,

2011). The three main phases observed following the latex-solvent interaction were solvent phase, gel phase and rubber coagulant. Sometimes, an aqueous phase was also observed due to water content in fresh latex. The carotenoid yield was obtained by recovering the organic phase and discarding the rubber coagulant and the gel phase. Therefore, the selection of the most suitable organic phase in this study was based on the yield (in volume) of solvent phase from the sample-solvent mixture.

The second part of the extraction optimisation was to evaluate the time needed for the optimal recovery of the extracted metabolites. This was determined through sequential solvent extraction of the liquid latex sample. The sequential extraction was performed by adding one part of liquid latex (volume/volume for liquid latex) to five parts of 100% acetone. The sample:solvent mixture was stirred continuously for 16 hours , in ambient temperature and under minimal light. Next, both coagulated rubber and solvent phase of the sample:solvent mixture were recovered. While the coagulated rubber was transferred into a new clean tube for the first sequential extraction, the recovered solvent phase was filtered through a Whatman No. 1 filter paper into a new and clean 15-ml tube. The solvent phase was dried under a gentle stream of nitrogen and kept in the absence of oxygen at -20^{0}C until further use. The extraction was repeated twice (hence the sequential extractions), on the same coagulated rubber, using a fresh batch of 5 ml acetone. Altogether, there were three dried extraction products generated from this optimisation.

The third part of the optimisation was to determine the recovery percentage of the spiked compound from both liquid latex and coagulated latex samples. The spiking of latex samples was achieved by adding trans-β-Apo-8′-carotenal in ethyl acetate into the liquid latex to a final concentration of 50

µg/ml. Next, the spiked latex was divided into two parts; the first part consisted of liquid spiked latex and the second part involved coagulating the spiked latex. The coagulation of the spiked latex samples was performed by adding 10 µl 100% ethanol to 400 µl of the spiked liquid latex. Both liquid and coagulated spiked latex were then subjected to solvent extraction. The solvent extraction was performed by making a sample:solvent mixture (1 part of liquid latex (volume/volume for liquid latex) to five parts of 100% acetone). The sample:solvent mixture was incubated for 16 hours, in the dark and afterwards, the solvent phase was recovered and dried using the drying step as previously described. A total of two extraction products (from spiked fresh latex and from coagulated spiked latex) were obtained from this parameter optimisation.

The extraction products from all optimisations were subjected to chromatographic separation (as detailed in the section 1.8). Concentration of the recovered spiked standard was calculated based on a 5-point calibration curve of known trans-β-Apo-8′-carotenal concentrations.

2.5. Extraction of isoprenoid intermediates from latex samples

For the extraction of isoprenoids from plant materials, two approaches reported by Li and Sharkey (2013) and Sajari et al. (2014) were used. Based on the technique by Li and Sharkey (2013), 200 mg of ground tissues of *Solanum tuberosum* and 200 µl fresh latex were added to 1 ml of acetonitrile-isopropanol-50 mM ammonium acetate (3:1:1, v/v ratio). The mixture was centrifuged at 5,000 g for 10 min and the supernatant was stored at -80 °C for later analysis.

To extract metabolites using the technique by Sajari et al. (2014), a 200 mg of ground *S. tuberosum* tissues and 200 µl fresh latex were added to 1 ml of chloroform-methanol solvent mixture (1:1, v/v ratio). The mixture was centrifuged at 5,000 g for 10 minutes and the supernatant was retained in a fresh 2 ml tube. The supernatant was subjected to a washing step, with 0.2 volume of 0.58% sodium chloride solution and incubated overnight at 4 ^{0}C. Afterwards, the mixture was centrifuged at 5,000 g for 20 minutes and the top layer was retained in a clean 1.5 ml tube. The top layer was dried under vacuum centrifugation for 1 hour. The dried samples were reconstituted in 200 µl deionised water:acetonitrile (1:1, v/v) prior to the chromatography run. For the extraction of isoprenoids from latex samples, 200 to 400 µl volume of liquid latex was mixed with 1 ml of chloroform-methanol solvent mixture (1:1, v/v ratio). The subsequent stage was performed according to steps described for the extraction of the potato samples.

2.6. Saponification of carotenoid extracts

Saponification is an approach often used to hydrolyse carotenoid esters and remove lipid components from the extraction product (Kimura et al., 1990, Larsen and Christensen, 2005). To perform saponification on the extracted carotenoids in this study, the dried latex carotenoid extracts were dissolved in 5 ml methanolic potassium hydroxide (10% weight/volume in 100% methanol). The reconstituted mixture was transferred into 20 ml quick-fit glass tubes and incubated overnight at 4 ^{0}C in the dark, with the absence of oxygen. Next, 10 ml of degassed diethyl ether with saturated sodium chloride was added, so that the mixture could be separated into two separate layers. The top layer was

removed into a fresh and clean 20 ml quick fit glass tube and washed with 10 ml deionised water. Finally, the top phase was transferred into a new 15 ml polypropylene tube and dried under a gentle stream of nitrogen.

2.7. Total carotenoid evaluation

The measurement of total carotenoid content was performed based on the absorbance of the extraction products at 450 nm using the U-3010 Spectrophotometer (Hitachi, Tokyo, Japan). The total carotenoid content was calculated using the extinction coefficient, *E* for mixtures of carotenoids at 2,500, based on a method reported by Schiedt and Liaan-Jensen (1995). The equation below was used to estimate the carotenoid amount, where *X* (g) weight of carotenoid in *y* (ml) volume of solvent gives and absorbance at A_{450} nm, with the assumption of a 1 cm path length of quartz cuvette

$$X = A_{450nm} \times y/(E \times 100)$$

2.8. High performance liquid chromatography separation of carotenoid extracts

Separation of latex carotenoids was performed using a high-performance liquid chromatography (HPLC) technique, as reported by (Fraser et al., 2000). The HPLC analytical method was implemented by using the Agilent 1260 Infinity Quarternary LC system (Agilent Technologies, USA). The instrument

was connected to the following instruments: Agilent 1260 Infinity Quaternary Pump (Agilent Technologies, USA), Agilent 1260 Infinity High performance Autosampler (Agilent Technologies, USA), Agilent 1260 Infinity Thermostated Column Compartment (Agilent Technologies, USA) and Agilent 1260 Infinity diode array detector (DAD) (Agilent Technologies, USA). The chromatographic separation was performed using YMC Carotenoid reverse-phase C30 column (250 x 4.6mm, 5µM) (YMC Inc, USA), fitted to YMC Carotenoid C30 guard cartridge (20 x 4.6mm) (YMC Inc, USA). The column temperature was maintained at 25 °C. The mobile phase used in this experiment consisted of: A, methanol, B, water/methanol (20:80, v/v) with ammonium acetate (0.1%, w/v), and C, tert-methyl butyl ether. The gradient of the mobile phase with a flow rate of 1 ml/minute used in this study is shown in Table 2.8.1.

Table 2.8.1: Mobile phase gradient used to separate carotenoid extracts from the latex samples

Chromatographic run time (min)	**Mobile phase solvents**		
	A	**B**	**C**
1.00 – 11.59	95%	5%	-
12.00-13.00	80%	5%	15%
13.01-43.00	30%	5%	65%
43.01-60.00	95%	5%	-

A= methanol; B= water/methanol (20:80, v/v) with ammonium acetate (0.1%, w/v); C= *tert*-methyl butyl ether

The carotenoid extraction products, together with the carotenoid standards were analysed in a single HPLC run. The carotenoid standards were prepared as serially diluted concentrations (between 0.39 µM – 100 µM) and a calibration curve was obtained based on the peak areas from the separated carotenoids. For the HPLC run, a total of 20 µl of sample was injected into the HPLC system and the carotenoids eluent was monitored by DAD. Under the DAD monitoring system, the samples were scanned through three discrete channels, namely at 286, 390 and 451 nm absorption wavelengths. Most of the

carotenoid compounds were eluted between the 10th to 35th minute during the 60-minute chromatographic separation. The data was analysed with the Agilent MassHunter Qualitative Software version B05.00 (Agilent Technologies, USA). The assignment of carotenoid identity was performed by manual integration of the chromatogram peaks and by comparing its retention time to that of the carotenoid standards. In addition to using carotenoid standards as a reference, the assignment of peak's identity was also supported by reported carotenoid spectral characteristics (Fraser et al., 2000). The amount of a given carotenoid was quantified by comparing its peak area to the calibration curve generated from the carotenoid standards.

2.9. Hydrophilic interaction liquid chromatography (HILIC) of isoprenoid intermediate extracts

Prior to utilising the HILIC protocol in identifying isoprenoid from the extraction products, few parameters were optimised. The parameters include i) the selection of a suitable HILIC column and ii) the determination of a working HILIC mobile phase. These parameters were evaluated based on the resolution of the seven isoprenoid intermediate standards generated in a single separation (Table 2.9.1). The working pipeline involves method development, isoprenoid intermediate separation and analytes interpretation using HILIC (as described sections 2.9 – 2.12). It was performed by Dr William Allwood of the Environment and Biochemistry Sciences group, the James Hutton Institute.

The first aspect of the method development was to select a suitable column to separate isoprenoid intermediates. Two different columns were used to separate the analytes (Table 2.9.2). The two columns provided different

techniques of analyte retention, due to their different stationary phase materials. The second aspect of the method development was to determine a working mobile phase by assessing the mobile phase's gradient (Table 2.9.3). An optimised mobile phase parameter assists in achieving a high resolution of HILIC peak for the corresponding analytes.

The HILIC run was operated using the Thermo Accela 600 HPLC system® coupled with an Accela PDA® detector and LTQ-Orbitrap XL mass spectrometry system® (Thermo-Fisher Ltd, U.K). The mobile phase consisted of 10 mM ammonium acetate pH 10 (solvent A) and acetonitrile (solvent B). The HILIC column was conditioned with solvents A and B (20:80; percentage; vol/vol) for a minimum of 80 mins at a flow rate of 200 µl/min. After the condition of the column, 20 µl of individual sample was injected into the HILIC column and the analyte was run at a flow rate of 200 µl/min.

Table 2.9.1: The isoprenoid intermediate standards used in the HILIC run. The information regarding monoisotopic mass and molecular formula was obtained from the corresponding material data sheet.

Isoprenoid intermediate	**Monoisotopic mass (Da)**	**Molecular formula**
Geranylgeranyl diphosphate (GGPP)	450.193627	$C_{20}H_{36}O_7P_2$
Farnesyl diphosphate (FPP)	382.131026	$C_{15}H_{28}O_7P_2$
Geranyl diphosphate (GPP)	314.068426	$C_{10}H_{20}O_7P_2$
Dimethylallyl diphosphate (DMAPP)	246.005826	$C_5H_9O_7P_2$
Isopentenyl diphosphate (IPP)	246.005826	$C_5H_9O_7P_2$
1-deoxy-d-xylulose 5-phosphate (DXP)	214.024239	$C_5H_{11}O_7P$
Methylerythritol phosphate (MEP)	216.039889	$C_5H_{13}O_7P$

Table 2.9.2: Columns used to separate isoprenoid analytes in the development of HILIC method for isoprenoid intermediate identification.

Column	Embedded particle	Functional group of the stationary phase	Dimension	Particle size (µm)
SeQuant® ZIC®-pHILIC	polymer	sulfobetaine-group	150x2.1 mm	5
Accucore 150 Amide HILIC	silica	amide	100x2.1 mm	2.6

Table 2.9.3. Gradients applied for the mobile phase of HILIC for the method development. The compositions of the mobile phase are 10 mM ammonium acetate (Solution A) and 100% acetonitrile (Solution B).

Gradient 1

Time (min)	Solution A (%)	Solution B (%)
0.00	20	80
2.00	20	80
2.01	30	70
6.50	30	70
7.50	60	40
10.00	60	40
10.01	20	80
15.00	20	80

Gradient 2

Time (min)	Solution A (%)	Solution B (%)
0.0	20	80
2.00	20	80
3	30	70
7.50	30	70
8.50	60	40
11.00	60	40
12	20	80
17	20	80

Gradient 3

Time (min)	Solution A (%)	Solution B (%)
0.0	20	80
4.00	20	80
6	30	70
15	30	70
17	60	40
22	60	40
24	20	80
34	20	80

2.10. Mass spectrometry (MS) of carotenoid compounds

Apart from acquiring data through the DAD system, information about the analyte was also obtained from multiple-reaction monitoring (MRM) in the Triple Quadrupole Mass Spectrometer system. For the MRM reaction, the analyte (from HPLC run) was ionised using the described MS settings together with the MRM transition and collision dissociation settings, which are summarised in Table 2.10.2.

To provide additional evidence for identification of the carotenoid peaks, MS/MS was performed using a Thermo LCQ Fleet Ion Trap Mass Spectrometer operated using the Xcalibur software package (Thermo Finnigan, San Jose, CA, USA). Firstly, 20 µl of sample was subjected to a reverse-phase chromatographic separation according to running conditions described in the previous section 2.8. Subsequently, MS of the separated compounds was performed using electrospray ionisation (ESI) in positive ionisation mode and the MS operating settings are summarised in Table 2.10.3. Intially, MS data was acquired through a full scan of the analytes. Afterwards, the top three most abundant ions revealed by the full MS scan were targeted in MS/MS scans (the scanning event conditions were set as follows: two *m/z* trapping widths, 25 millisecond activation time, 35% normalised collision energy, helium as collision gas, MS scan speed 1:10 milliseconds).

Table 2.10.1: The mass spectrometer operating settings used in Agilent 6460A Triple Quadrupole Mass Spectrometer for carotenoid compound ionisation

Operating settings	**Detail**
Drying gas temperature	300 °C
Drying gas flow rate	5 L/minute
Nebulizer pressure	45 psi
Sheath gas temperature	250 °C
Sheath gas flow rate	11 L/minute
Capillary cap voltage	4.5 kV
Nozzle voltage	500 V

Table 2.10.2: The transition ion settings used for multiple-reaction monitoring of carotenoid compounds

Analyte	**Ion species**	**Precursor *m/z***	**Product *m/z***	**Fragmentor voltage**	**Collision energy**
Lutein	$[M]^{+\bullet}$	568	476, 119, 105	84	8
Zeaxanthin	$[M]^{+\bullet}$	568	476, 119, 105	96	8
β-cryptoxanthin	$[M+H]^{+}$	553	461, 119, 105	99	8
α-carotene	$[M]^{+\bullet}$	536	444, 119, 205	98	12
β-carotene	$[M]^{+\bullet}$	536	444, 119, 205	98	8

Table 2.10.3: The running condition settings used in the ionisation of carotenoid compounds in the LCQ Fleet Ion Trap Mass Spectrometer

Operating settings (LCQ Fleet Ion Trap Mass Spectrometer)	**Detail**
ESI heater temperature	150 ^{0}C
Sweep Gas	0 AU
Sheath gas flow	60 AU
Auxiliary gas flow	20 AU
Spray voltage	+4 kV /-3.5 kV
Ion tube temperature	280 ^{0}C
Ion tube voltage	+10 / -10 V

2.11. Mass spectrometry of isoprenoid intermediates

The analytes eluted from the HILIC technique (performed as previously described in section 2.9) were channelled to the Thermo LTQ-Orbitrap XL mass spectrometry system operated by the Xcalibur software (Thermo-Fisher Ltd. U.K.) for MS. First, the analytes were ionised using electrospray ionisation (ESI) in negative mode. The following settings were applied to the ESI process: spray voltage at -3.5kV; 35 and 15 arbitrary units of sheath gas and auxiliary gas; capillary voltage at 35V; 100 V tube lens voltage; and capillary temperature was set at 380 °C. Then, the ionised analytes were subject to full MS scan analysis. The full MS scan data was acquired within the Fourier Transform (FT), with a detection range of *m/z* 70-1000 *m/z*. Based on the full MS scan data, three most abundant ions were targeted for the subsequent MS/MS scan. The MS/MS scan was applied through the collision ionisation dissociation (CID) of the targeted ions. Data from both full scan MS and MS/MS methods was collected centroid mode.

For HILIC run of extracted samples, the LC-MS system was initially conditioned with eight injections of the sample mixture (known as quality assurance or QA). The QA was generated by mixing an equal amount of individual sample and analytical controls into a single tube. The data generated from QA of HILIC-MS/MS would give an indication on the LC-MS system stability.

2.12. HILIC-MS/MS data analysis

The HILIC-MS/MS raw data profiles were firstly converted into an MZML centroid format using the Proteowizard MSConvert software (http://proteowizard.sourceforge.net/). Each MZML-format data was converted into extracted ion content (EIC) peak areas, where a peak response was defined as the sum of intensities over a window of specified mass and time range (e.g. m/z = 102.1 ± 0.01 and time = 130 ± 30 s). The peak area extraction was performed using the XC-MS software (http://masspec.scripps.edu/xcms/xcms.php).

Next, normalisation of the extracted peak area of an individual sample was performed by comparison with the EIC of the QA data (QA data (which was generated from the mixing of equal amounts of individual sample and analytical controls). The peak area of an individual sample exhibiting relative standard deviation (RSD) that was higher than that of the QA peak area would be not be included in the subsequent data analysis.

For putative metabolite identification, the *m/z* of individual peaks were compared against the PutMedID database in the Taverna Workbench 1.7.2 software (Wolstencroft et al., 2013). In addition, the ion species formed during the ionisation process were also defined based on the mass differences between correlated *m/z* features with the common ion adducts in ESI negative mode such as H^-, Cl^- or $CHOOH^-$. This permitted the calculation of the compounds' neutral mass from the accurate mass MS information for the charged ion. Once the neutral mass was known, it was matched to the possible molecular formula(s) and metabolite names for each molecular formula through comparison to the Manchester Metabolite Database (Brown et al., 2009). In

addition to the accurate mass data annotation, identification was also made though comparison of the individual peaks to that of the corresponding analytical standards.

2.13. Dry rubber content measurement

The measurement of dry rubber content (DRC) was performed on fresh latex collected from PB235 and RRIM600 *Hevea* genotypes (as previously mentioned in section 2.1). The measurement was carried-out based on a published generic protocol used to calculate DRC from field latex (Rubber Research Institute of Malaysia, 1973, Chin and Singh, 1980) with the assistance from Mr Ahmad Khairul and Ms Noor Hiza. Briefly, 10 ml of the individual latex sample was poured into a petri dish. Coagulation of the latex sample was initiated by slowly adding 150 ml of 0.5% acetic acid (vol/vol) (under constant stirring). The coagulated latex was rolled into a thin sheet on a clean aluminium plate. The coagulated latex and smaller particles of coagulated rubber particles on the aluminium plate were collected and dried overnight in 70 °C. The dried coagulated latex was weighed, and the DRC was expressed as follows:

$$\text{DRC (\%)} = \frac{\text{weight of dry coagulated latex}}{\text{weight of latex sample}} \times 100$$

2.14. Latex collection and total RNA extraction

Total RNA extraction from latex samples was performed based on the method described by Kush et al. (1990). Latex samples were obtained from the *Hevea* tree genotypes described in section 2.1. Similarly, the latex collection

was performed according to the protocol described in section 2.1. After the tapping process, exuded latex was allowed to flow for 10-15 seconds, to exclude debris on the tapping groove and spout. After excluding the initial exudate, 20ml latex was then collected directly into 50 ml Falcon tubes containing 20 ml of 2X RNA extraction buffer (0.1 M Tris-HCl, 0.3 M LiCl, 10 mM EDTA, 10% SDS, pH 9.5; the pH was adjusted with 10 M NaOH). Due to the propensity of fresh latex to coagulate, the mixing of the collected latex into RNA buffer extraction was performed *in situ.* As the latex dripped into the buffer, constant stirring with a sterile glass rod was applied. The mixture of latex:buffer was placed on ice. The latex:buffer mixture was centrifuged at 59,000 g for 15 minutes at 4 ^{0}C and the serum fraction was recovered with a glass pipette and transferred into a 50 ml polypropylene centrifuge tube. Next, an equal volume of phenol/chloroform buffer (1:1 vol/vol, pH 7.4) was added to the recovered latex serum. This mixture was mixed by inverting the centrifuge tube several times and then centrifuged at 59,000 g for 15 minutes at 4 ^{0}C. The resulting supernatant (aqueous phase) was transferred into a new 50 ml polypropylene centrifuge tube. Thereafter, chloroform extraction was performed by adding an equal volume of chloroform to the supernatant. The tube containing the supernatant:chloroform mixture was inverted several times and then centrifuged at 59,000 g for 15 minutes at 4 ^{0}C. The supernatant (aqueous phase) was transferred into a new 50 ml polypropylene centrifuge tube and subjected to RNA precipitation. The precipitation was performed by adding 8 M $LiCl_2$ stock solution into the supernatant, to achieve a final $LiCl_2$ concentration of 2 M. After an overnight incubation at-80 ^{0}C, the mixture was centrifuged at 59,000 g for 20 minutes at 4 ^{0}C to pellet latex total RNA. Finally, the supernatant was

discarded, and the total RNA pellet was reconstituted in 100 µl cold sterile water.

2.15. RNA quality assessment

RNA concentration and quality were assessed using the NanoDrop® ND-1000 Spectrophotometer (Thermo Fisher Scientific Inc, USA). RNA concentration was obtained by measuring sample absorbance at 260 nm and RNA quality was determined from the ratio of sample absorbance values at wavelengths 230, 260 and 280 nm. The isolated total RNA was also subjected to RNA integrity number assessment using the Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and the RNA 6000 Nano LabChip kit (Caliper Technologies, USA). The RNA Integrity Number (RIN), fragment size distribution, concentration and rRNA ratios were generated by the Agilent 2100 Bioanalyzer system software version B.01.03. An RNA integrity number (RIN) greater than eight is recommended for all samples (Schroeder et al., 2006).

2.16. RNA clean-up

Some RNA samples were found to have organic solvent contamination, as reflected by the low spectrometry absorbance ratio at 260 and 230 nm. The removal of such contaminants was performed through the extraction of the latex total RNA using a solvent mixture consisting of phenol, chloroform and isopropanol (25:24:1 ratio, vol/vol/vol). A total of 400 µl total RNA was added to an equal volume of the solvent mixture in a sterile 1.5 ml tube. The mixture was

centrifuged at 18,500 g for 5 minutes using a benchtop high speed centrifuge (Eppendorf 5424, Eppendorf UK). The upper aqueous phase was transferred into a new sterile 1.5 ml tube. Next, RNA extraction was performed using a solvent mixture consisting chloroform and isopropanol (24:1 vol/vol). The mixture was centrifuged at 18,500 g for 5 minutes. The aqueous phase was recovered into a new 1.5 ml tube. A total of 0.1 volume (of the transferred aqueous phase) of 3M sodium acetate pH 5.2 and 2.5 volume (of the transferred aqueous phase) of absolute ethanol were added into the recovered aqueous phase. The tube containing the mixture was inverted a few times and then incubated overnight at -80 ^{0}C for 12 hours. Following the overnight incubation, the mixture was centrifuged at 18,500 g for 5 minutes to obtain RNA pellet. The supernatant was removed and ethanol solution (70%; vol/vol) was added to the pellet for washing step. The washing step was performed by mixing the pellet with the ethanol solution thoroughly, followed by centrifugation at 18,500 g for 5 minutes. Then, the supernatant was removed, and the pellet was air-dried before being re-dissolved in 50 µl sterile deionized water.

2.17.1. cDNA library construction and preparation for sequencing

The work pipeline starting from the cDNA library construction from the latex total RNA, leading to the pooling of the libraries (sections 2.17.1 – 2.17.8). It was performed by Mrs Jenny Morris from the Genome Technology group, the James Hutton Institute. Active observations and interactions with Mrs Morris were made during the process of the library construction. Final decisions regarding the experiments were made based on suggestions by Mrs Morris.

The RNA-seq cDNA libraries were constructed from total RNA samples extracted from *Hevea* lines PB235 and RRIM600 (6 replicates for each line). The library construction was performed in 96-well plates using TruSeq® RNA Sample Preparation Kit version 2 (Illumina, San Diego, USA), based on the Low Sample Protocol, according to the vendor's instruction. All PCR components, ligation reaction components, reverse transcription components, elution and washing buffers were provided with the kit. Briefly, for the library construction, mRNA was captured by using oligo-dT. Subsequently, the captured mRNA is fragmented using divalent cations under elevated temperature (94 °C, 8 minutes). The fragmentation process ended up with the mRNA fragments ranging from 125 – 250 bp, with a median size 165 bp. The fragmented mRNA was reverse-transcribed into cDNA using random hexamers. The cDNA fragments were subjected through an end repair process, the addition of a single 'A' base, and then ligation of the adapters as summarised in the eight key steps illustrated in Figure 2.17.1.1. Rather than cloned into plasmid vector, the purified and enriched cDNA fragments in each library were directly loaded onto the surface of the sequencing flow-cell (Illumina) for the cluster generation. The following sections described in greater details the steps to generate the final library prior to loaded onto the sequencing flow-cell.

For the purification and fragmentation process, 1 µg of latex total RNA was first diluted with nuclease-free ultra-pure water to a final volume of 50 µl. The diluted RNA was added to 50 µl oligo-dT beads (provided as Bead Binding Buffer) in a sterile 96-well PCR plate and the mixture was mixed thoroughly by repeated pipetting. The mixture was incubated at 65^0C for 5 minutes, followed by incubation at room temperature for 5 minutes. To separate the polyA-RNA bound beads from the mixture solution, the PCR plate was placed on a

magnetic stand and left for 5 minutes at room temperature. The mixture would separate into supernatant and bead pellet layers after the five-minute incubation. The supernatant was removed and the pellet was reconstituted in 200 µl Bead Washing Buffer. The washing step was performed by gently pipetting the mixture. The mixture was incubated at room temperature for five minutes before being incubated on the magnetic stand for 5 minutes, under room temperature. The mixture would be separated into supernatant and bead pellet layers after the 5-minute incubation. The supernatant was discarded and the pellet was reconstituted in 50 µl Elution Buffer. Then, the mixture was incubated at 80 ^{0}C for 2 minutes, for the removal of the oligodT beads from the mRNA. A total of 50 µl Bead Binding Buffer was added to the PCR plate to allow mRNA to rebind to the new beads. The supernatant was discarded from the PCR plate. Finally, the washing step was performed using 200 µl Bead Washing Buffer, as described in the above step. The washed beads were used in the subsequent first-strand cDNA synthesis.

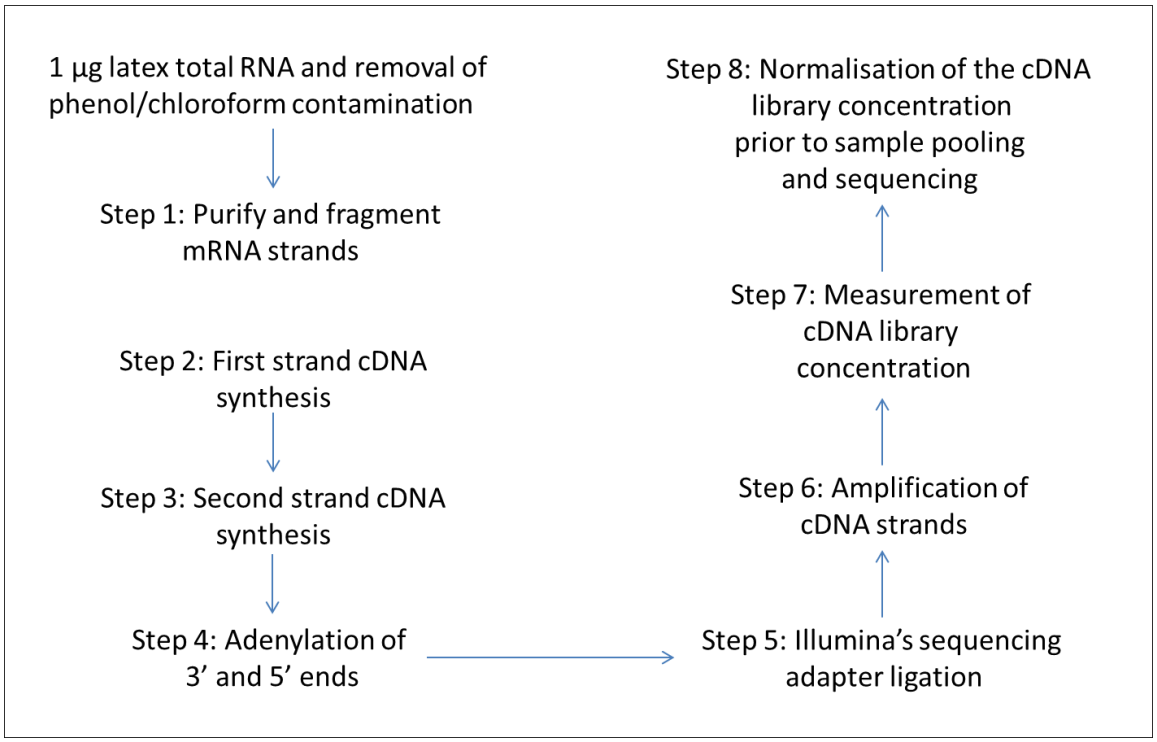


Figure 2.17.1.1: Flow chart illustrating the steps of cDNA library preparation prior to RNA sequencing. The cDNA library construction was performed according to TruSeq RNA Sample Preparation v2 (Illumina, USA) to generate cDNA inserts ligated to sequencing adapters with lengths between 200 – 500 bp.

2.17.2. First-strand cDNA synthesis

A total of 19.5 µl of Elute, Prime, Fragment Mix solution was added to the washed beads and the mixture was pipetted to mix thoroughly. The Elute, Prime, Fragment Mix contained random hexamers for first strand cDNA synthesis. This allowed the primers to anneal to the purified mRNA (that was bound to the beads). The mixture was incubated at 94 ^{0}C for 8 minutes. Next, the plate was incubated on a magnetic stand at room temperature for 5 minutes to allow the separation of the mixture into supernatant and pellet phases. Afterwards, 17 µl of the supernatant was transferred into a new sterile 96-well PCR plate and this served as mRNA template for the subsequent first strand cDNA synthesis. Then, the mRNA template was combined with 6.25 µl First Strand Master Mix and 1.25 µl SuperScript II reverse transcriptase for the reverse transcription reaction. The first-strand cDNA synthesis was performed using the PCR cycle profile in Table 2.17.2.1.

Table 2.17.2.1: PCR profiles used during conversation of total RNA into first-strand cDNA

PCR step	**Temperature (^{0}C)**	**Time (seconds)**
Annealing	25	10
Polymerisation	42	42
Reaction termination	70	70

2.17.3. Second-strand cDNA synthesis

The synthesis of second-strand cDNAs was performed by adding 25 µl Second-strand Master mix to the first strand cDNA products. The mixture was mixed thoroughly and incubated at 16 ^{0}C for 60 minutes. Afterwards, the second-strand cDNAs were isolated from the remaining reverse transcription

components through a purification step using AMPure XP beads (Illumina, US). The purification was performed by binding the DNA to the beads, followed by two rounds of washing steps using 80% (vol/vol) ethanol and elution of the bound DNA from the beads. For the binding of DNA to the beads, 90 µl AMPure XP beads and the mixture was incubated at room temperature for 15 minutes. The PCR plate (containing the cDNA products bound to the beads) was then placed on a magnetic stand for 5 minutes to separate the DNA-bound beads from the aqueous phase. Whilst the PCR plate was still placed on the magnetic strand, the resulting supernatant was discarded and the pellet (DNA-bound beads) was washed with 80% ethanol (vol/vol). The washing step was performed by adding 200 µl 80% ethanol (vol/vol) and the mixture was incubated for 30 seconds prior to removing the supernatant without disturbing the pellet. After the washing step was performed twice, the pellet was air-dried at room temperature for 15 minutes prior to adding 52.5 µl Resuspension Buffer. By adding Resuspension Buffer, the bound DNAs were released from the magnetic beads. After the mixture was incubated for 2 minutes in room temperature, the PCR plate was placed on the magnetic stand for five minutes. By placing the PCR plate on the magnetic stand, the mixture was separated into supernatant and magnetic beads pellet. Fifty µl of the supernatant was transferred to the new 96-well PCR plate for cDNA ends repair.

2.17.4. Adenylation of cDNA ends

Prior to the ligation of adenosine nucleotide to the cDNA ends, sticky ends of the cDNA strands were converted to blunt-ends through a cDNA repair step. The cDNA ends repair was performed by adding 10 µl Resuspension

Buffer to the 50 µl of the supernatant isolated at the end of section 2.17.4). Then, 40 µl End Repair Mix was added to the mixture and incubated at 30 ^{0}C for 30 minutes. Next, the cDNA with repaired ends were purified using 160 µl AMPure XP (Illumina, USA) based on the purification steps previously described in section 2.17.4.

A total of 15 µl of purified cDNAs with repaired ends were transferred into a new 96-well PCR plate and were subjected to the ligation of adenosine nucleotide to the cDNA ends. The cDNA ends adenylation was carried-out by adding 2.5 µl Resuspension Buffer (Illumina, USA) and 12.5 µl A-Tailing Mix (Illumina, US) to the purified cDNAs. The ligation step was performed by incubating the mixture at 37 ^{0}C for 30 minutes, followed by an incubation step at 70 ^{0}C for 5 minutes. Afterwards, the ligation products were subjected to another ligation step to add multiple indexing adapters to the cDNA ends. A total of 12 different indices were used as listed in (Table 2.17.4.1). The ligation process was carried out by adding 2.5 µl Ligation Mix (Illumina, San Diego, US) and 2.5 µl RNA Adapter Index (Illumina, San Diego, US) to the cDNA samples. The mixture was incubated at 30 ^{0}C for 10 minutes. To stop the reaction, 5 µl Ligation Buffer was added to the ligation mixture reaction. Afterwards, purification of the cDNA with overhanging ends were performed by adding 42 µl AMPure XP beads to the ligation mixture reaction. The purification was performed according to the steps previously described in section 2.17.4. A total of 20 µl of the purified ligated cDNAs were transferred into a new 96-well PCR plate. The cDNAs would serve as template for the subsequent amplification step.

Table 2.17.4.1: Multiplexing index adapters used for libraries generated from latex total RNA of RRIM600 and PB235 *Hevea* genotypes. The number at the end of each genotype indicate the identifier of the individual library.

Library sample	Index	Index sequence
RRIM600.3	AR001	ATCACG
RRIM600.4	AR003	TTAGGC
RRIM600.5	AR008	ACTTGA
RRIM600.6	AR009	GATCAG
RRIM600.8	AR010	TAGCTT
RRIM600.9	AR011	GGCTAC
PB235.3	AR020	GTGGCC
PB235.4	AR021	GTTTCG
PB235.5	AR022	CGTACG
PB235.6	AR023	GAGTGG
PB235.7	AR025	ACTGAT
PB235.8	AR027	ATTCCT

2.17.5. PCR amplification of cDNA samples

For cDNA amplification, 5 µl PCR Primer Cocktail and 25 µl PCR Master Mix were added to the cDNA template and the mixture was subjected to the PCR cycles described in Table 2.17.5.1. Then, the PCR products were purified using AMPure XP beads as previously mentioned in section 2.17.4. After the purification was completed, 30 µl of the clean amplified cDNA were transferred into a new 96-well PCR plate and kept at -80 °C until further use.

Table 2.17.5.1: PCR profile for the amplification of cDNA ligated to multiplex index adapters

PCR step	Temperature (^{0}C)	Time (seconds)	Cycle number
Denaturation	98	30	1
Annealing	60	10	
Polymerisation	72	30	15
Final polymerisation	72	300	1

2.17.6. Measurement of library concentration

The measurement of library concentration was performed using a combination of three different approaches, namely electrophoresis through the Agilent DNA 1000 Bioanalyzer (Agilent Technologies, US), spectrometry absorbance assay using the Qubit® 2.0 Fluorometer (Thermo Fisher Scientific, US) and amplification of the library by quantitative polymerase chain reaction (qPCR) in StepOne™ Real-Time PCR (Thermo-Fisher, USA). This would generate information related to insert size distribution, purity and concentration of the libraries. Prior to the assessment assay, the libraries were diluted (1:1, vol/vol) with Resuspension Buffer. The diluted cDNA library was loaded onto a DNA-specific chip known as the Agilent DNA 1000 chip (Agilent Technologies, US) and electrophoresis was performed according to the vendor's manual.

Spectrometric absorbance of the double stranded DNA contained inside the libraries was measured using Qubit® dsDNA BR Assay Kit (Broad range DNA assay) (Thermo Fisher Scientific, US), in conjunction with Qubit® 2.0 Fluorometer. The assay was carried out according to the manufacturer's instructions. Prior to measuring the DNA concentration, a working solution buffer was prepared by diluting the Qubit® dsDNA BR Reagent to 1:200 (vol/vol) in Qubit® dsDNA BR Buffer. Likewise, the constructed libraries were diluted to 2:200 (vol/vol) in sterile deionised water. Two DNA standards, Qubit® dsDNA BR Standard #1 (0 ng/µl) and Qubit® dsDNA BR Standard #2 (100 ng/µl) as provided in the Qubit® dsDNA BR Assay Kit, were utilised in this assay. Each assay reaction tube was set up by adding together 190 µl of working solution buffer and 10 µl of diluted libraries or DNA standard. The reaction tubes were incubated at room temperature for 2 minutes before measuring the absorbance. A standard curve based on the two standards was generated based on a curve-fitting algorithm and was used in the calculation of the libraries' concentration using the acquired Qubit data. The spectrometry measurement was performed based on the Qubit® 2.0 Fluorometer User Guide (Thermo Fisher Scientific, US). The concentration of each libraries was calculated based on the following equation:

$$\text{Library concentration} = \text{QF value}^{*} \times \frac{200}{x^{\#}}$$

* QF value refers to the reading value given by the Qubit® 2.0 Fluorometer
\# *x* refers to the volume of sample (in µl) used for the spectrometry assay

The measurement of DNA concentration through real time quantitative PCR (qPCR) would enable absolute quantity of DNA that contained Illumina adapter sequences on 3' and 5' –ends. The qPCR amplification was performed

using NEBNext® Library Quant Kit for Illumina® (New England Biolabs Inc, US), according the manufacturer's instructions. Prior to setting up the qPCR reaction, 1 µg of the library was serially diluted into 1:1000, 1:10,000 and 1:100,000 with 1X NEBNext Library Quant Dilution Buffer. The qPCR reaction was set up in a 0.3 ml 96-well PCR plate by mixing 15 µl NEBNext Quant Master Mix, 1 µl 20X NEBNext Library Quant Primer Mix and 4 ul of the diluted DNA. The amplification was performed using the PCR profile shown in Table 2.17.6.1 in the StepOne™ Real-Time PCR thermal cycler (Thermo Fisher Scientific, US). For each library, three qPCR runs were carried out. DNA standard of known amount (provided with the kit) was included in the same experimental condition (in triplicates) for the standard curve generation. The acquired data was viewed using StepOne™ and StepOnePlus™ Real-Time PCR System software (Thermo Fisher Scientific, US). The standard curve was generated by plotting signals of SYBR/FAM of the DNA standards against the corresponding DNA amount. The quantification of amplified cDNA was performed by comparing the Sybr/FAM signals of the amplified diluted libraries to the standard curve.

Table 2.17.6.1: PCR profile for the real time quantitative PCR for library concentration measurement

PCR step	**Temperature (^{0}C)**	**Time (seconds)**	**Cycle number**
Initial denaturation	95	60	1
Denaturation	95	15	35
Annealing/extension	63	45	

2.17.7. Normalisation of cDNA library and pooling into a single sample

After the assessment of both quantity and quality of the libraries, the libraries for PB235 and RRIM600 *Hevea* tree genotypes were subjected to the normalisation and pooling process. The normalisation of each library was performed by diluting the library to the final concentration of 10 nM in Tris-HCl 10 mM, pH 8.5 with 0.1% Tween 20. The volume needed for each library during normalisation was determined based on the quantity values generated from qPCR data. To combine the libraries, 24 µl each of the normalised libraries was pooled together and the pooled libraries were pipetted up and down 10 times to mix thoroughly.

2.18. Mi-Seq sequencing

Prior to sequencing with Illumina NextSeq, the libraries were subjected to Mi-Seq sequencing. The Mi-Seq sequencing reaction and raw data conversion was performed by Mrs Jenny Morris from the Genome Technology group, the James Hutton Institute. This was another approach to assess the quality of the constructed libraries. The sequencing was carried out using MiSeq® Reagent Kit v3 (Illumina, US) and the Illumina Mi-Seq® system sequencer (Illumina, US). Briefly, 20 nM of the pooled libraries were denatured with 0.2 N sodium hydroxide. The denatured library pool was further diluted to 15 pM final amount with Hybridisation Buffer 1, provided in the kit. The library was combined with 20 pM PhiX control DNA prior being loaded in the reaction cartridge (MS3813393-600 v3, Illumina, US). The paired-end reads sequencing was performed in a

single lane. Raw Mi-Seq data were converted into fastq-formatted files using RTA software RTA v1.18.5.4 (Illumina, US).

2.19. RNA sequencing

The pooled multiplexed libraries were sent to the Tayside Centre for Genomic Analysis at Ninewells Hospital, University of Dundee, UK for next generation sequencing and generation of the short-read data in fastq format. The sequencing of the pooled multiplexed libraries was performed using the Illumina NextSeq 500 sequencing system (Illumina, US). For the sequencing reaction, the pooled multiplexed libraries were loaded onto four lanes of Illumina's flow cell. An equal amount of the normalised multiplexed libraries was distributed into these four lanes. The de-multiplexing of the pooled libraries was performed by the sequencing provider based on the multiplexing index adapters. The sequencing reaction generated 150 bp paired-end reads for the libraries. The preparation and loading of the library and DNA control (Phi-X, Illumina, US) onto the reaction cartridge was performed using NextSeq 500 Kit (150 cycles) (Illumina, US), based on the vendor's instructions. The conversion of the primary sequencing data to fastq format was carried out using the bcl2fastq software (Illumina, USA).

2.20. Public transcriptome data

Publicly available RNA-seq data and draft genome sequence used for the construction of the reference transcript were downloaded from the National Center for Biotechnology Information (NCBI)'s short read archive (SRA) (https://www.ncbi.nlm.nih.gov/sra). The Iso-seq data generated from

Pootakham et al. (2017) was retrieved from the *Hevea* tree genome website (http://www4a.biotec.or.th/rubber/GenomeSeq). The full length cDNA data from *Hevea* tree genotype RRIM600 (Makita et al., 2017) was downloaded from RIKEN BioResource Center (BRC) (http://matsui-lab.riken.jp/rubber/). The full list of the downloaded data is described in Table 2.20.1.

Table 2.20.1: The public transcriptome raw data downloaded for transcript construction.

	***Hevea* tree genotype**	**Tissue**	**NCBI accession no.**	**Data size (gb)**
RNA-seq	Reyan 7-33-97	primary laticifer	SRR5118400	3.3
		primary laticifer	SRR5118398	2.4
		primary laticifer	SRR5118396	1.6
		secondary laticifer	SRR5118395	3.1
		secondary laticifer	SRR5118397	2.8
		secondary laticifer	SRR5 118399	1.8
		root	SRR3136156	3.2
		bark	SRR3136158	3.1
		leaf	SRR3136159	3.4
		latex	SRR3136162	2.0
		female flower	SRR3136165	4.0
		male flower	SRR3136166	3.6
		seed	SRR3136168	3.4
	RRIM600	petiole	DRR069093	12.1
		leaf	DRR069092	11.6
		latex	DRR069091	11.5
		bark	DRR069090	11.3
	PR 107	bark	SRR2147074	3.6
		bark	SRR2147073	3.6
		bark	SRR2147072	3.7
		leaf	SRR3240371	4.0
		latex	SRR2001614	1.4
		bark	SRR1611790	2.9
		bark	SRR15 88170	8.7
		latex	SRR15 33844	1.4
	RRIM 928	leaf	SRR854522	2.6
		bark	SRR854520	9.6
		latex	SRR854521	3.1
	BPM 24	latex	SRR1649353	8.9
	RRIC 110	latex	SRR1649349	13.8
	RRII 105	latex	SRR164935 0	11.2
	RRIM 901	latex	DRR069094	10.8

Iso-seq	BPM 24	leaf	*Not applicable	0.7
Draft genome sequence	Reyan 7-33-97	leaf	LVXX01000000	1.4
Full-length cDNA	RRIM600	leaf	*Not applicable	0.02

*Not applicable indicates the data does not have NCBI accession number. The data was downloaded from websites provided by their authors.

2.21. RNA-seq data processing

Transcript construction using RNA-seq reads was performed according to the steps illustrated in Figure 2.21.1. Initial processing steps namely removal of adapters and low-quality reads, mapping of the reads to the reference sequence, transcript assembly, and transcript merging were carried out. For the removal of adapters and low-quality reads and the mapping of the reads to the reference sequence steps, optimisation had to be performed to ensure minimal loss of data was incurred and only high-quality reads were retained. Finally, evaluation of the merged transcripts was performed to assess the completeness of the transcripts. All steps in RNA-seq data processing were described in the following sections (Sections 2.22 – 2.29).

2.22. RNA-seq quality evaluation and adapter trimming

The first step in RNA-seq pre-processing was to evaluate basic metrics that represent the quality of RNA-seq reads. Metrics include number of reads, distribution of base-call score, length of reads, nucleotide distribution, GC content, sequence duplicates and sequence complexity were evaluated. These metrics were viewed using FastQC software (Andrews, 2015). Next, the filtering-out of low-quality base from the reads was carried out. The quality of RNA-seq reads was encoded by an algorithm known as Phred. The Phred algorithm was developed to indicate the probability that the base is called incorrectly by the sequencer (Richterich, 1998). The Phred scores ranged from 0 (the least reliable base-call) to 40 (the lowest chance of error base-call).

Ambiguous bases (N) or bases with Phred quality score (Q score) less than 15 were discarded from the RNA-seq reads. In addition, the sequencing adapters in the RNA-seq reads were also removed. Both quality trimming and adapter filtering were performed using Trimmomatic software (Bolger et al., 2014). Only RNA-seq reads free of adapters and having Q score >15 and read length >50 bp were retained.

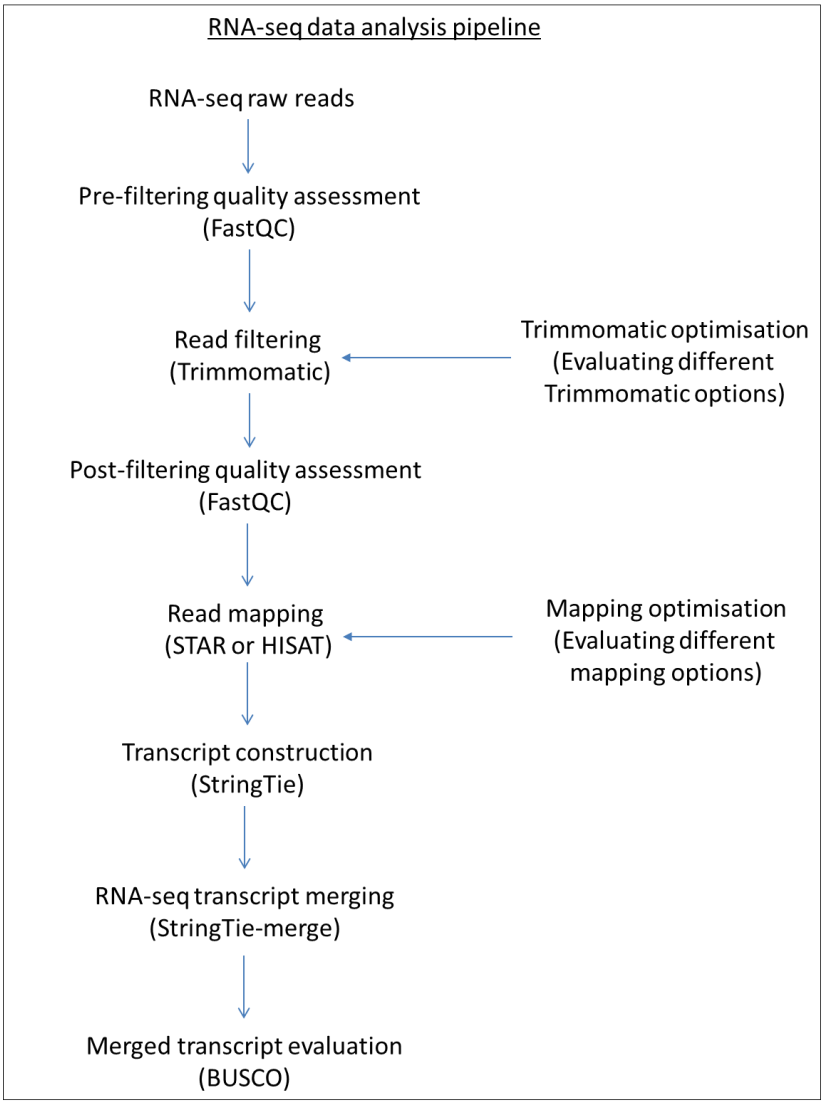


Figure 2.21.1: Pipeline of RNA-seq data analysis. The raw reads were cleaned by removing the adapter sequences and nucleotides with low Qual scores. Transcript construction was performed by mapping the cleaned reads to the sequence reference (draft genome sequence). The transcript sequences were merged, and evaluation of its completeness was performed based on the set of transcripts reported to encode housekeeping genes that are expected to be present in plants.

2.23. RNA-seq read mapping

Optimisation for the mapping quality was carried out so that the most suitable options for *Hevea* RNA-seq mapping could be obtained. Due to the large volume of the RNA-seq data, the optimisation was only performed on seven RNA-seq data set. The best parameters determined during the optimisation process were then applied for all RNA-seq involved in this study.

During the read mapping step, the intron length for the mapped reads was set between 60 bp to 6000 bp. This value was selected based on reported typical intron length in *Arabidopsis* (Marquez et al., 2012) and based on the advice by Dr Runxuan Zhang (personal communication, 2017). The parameters evaluated involved: i) software used for read mapping (STAR and HISAT); and ii) different mismatch rates for the mapped reads. The parameters set for mapping using STAR (Dobin et al., 2013) are listed in Table 2.23.1 and the parameters set for mapping using HISAT2 (Kim et al, 2016) are displayed in Table 2.23.2. Three different mismatch rates were used: no mismatch (0MM), 2 mismatches per mapped read pair (2MM) and 4 mismatches per mapped read pair (4MM).

Quality evaluation of the mapping step was performed by counting: i) reads that were mapped concordantly 1 time to the reference (known as unique reads), ii) reads which were mapped concordantly to multiple locations (known as multi-mapped reads) and ii) unmapped reads. The read count was performed using the featureCounts software (Liao et al., 2014).

Table 2.23.1: Options of STAR algorithm used for the mapping of RNA-seq reads. The options have been optimised and the evaluation process is described in Appendix section 2.2

STAR alignment	**Selected option**
--outSAMtype	BAM SortedByCoordinate
--outSAMstrandField	intronMotif
--outSAMprimaryFlag	AllBestScore
--outSJfilterReads	All
--outSJfilterCountUniqueMin	3 -1 1 1
--outSJfilterCountTotalMin	10 7 7 7
--outSJfilterOverhangMin	30 12 12 12
--outSJfilterDistToOtherSJmin	10 0 5 10
--outSJfilterIntronMaxVsReadN	50000 100000 200000
--outFilterType	BySJout
--outFilterMatchNminOverLread	Range between 0.96 to 1.00
--outFilterMismatchNoverReadLmax	Range between 0 to 0.04
--outFilterIntronMotifs	RemoveNoncanonical
--alignIntronMin	60
--alignIntronMax	6000
--alignSJoverhangMin	7
--alignSJDBoverhangMin	7
--sjdbOverhang	99
--twopassMode	Basic

Table 2.23.2: Options of HISAT2 algorithm used for the mapping of RNA-seq reads.

HISAT2 alignment	**Selected option**
--min-intronlen	60
--max-intronlen	6000
--score-min	L,0, range between -0.24 to 0
--no-unal	yes
--n-ceil	L,1,0

2.24. Transcript assembly

The assembly of transcripts was carried out using the StringTie software (Pertea et al., 2015). The assembly was performed based on two different parameters of the mapping software: i) minimum reads per bp coverage (known as ‘c’ parameter in the software algorithm) and ii) gap between read mappings that differentiating two different transcript bundle (known as ‘g’ parameter in the software algorithm).

2.25. Transcript merging

The transcripts assembled from the RNA-seq data sets were merged using the StringTie-merge algorithm (Pertea et al., 2015). The merging was performed using the default parameters suggested in the software manual.

2.26 The evaluation of transcript completeness

Completeness of the merged transcripts were evaluated using BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simão et al., 2015). The completeness of a transcriptome data is determined by evaluating a set of essential single-copy orthologs that are expected to be present in any plant kingdom against the assembled transcriptome. A qualitative measure is generated when the BUSCO software identifies the single-copy orthologs as complete, duplicated, fragmented and missing genes. A transcriptome with a high completeness will show a higher number of complete single-copy orthologs. The evaluation was performed with default settings of the software.

2.27. Error-correction of long reads

The Iso-seq data reported by Pootakham et al. (2017) comprised of circular consensus sequences (CCS) transcripts was corrected with RNA-seq. Although the Iso-Seq data has been processed to remove any adapter sequences, errors in the form of insertions and deletions were still present in the data (Weirather et al., 2017). To correct the errors, RNA-seq data originated from the same *Hevea* tree genotype was mapped to the Iso-seq data in a step known as hybrid assembly. Two different hybrid assembly software tools; proovread (Hackl et al., 2014) and LoRDEC (Salmela and Rivals, 2014) were evaluated to determine the best tool that generate the most amount of corrected sequence and the least number of the discarded reads. Default settings were used for each proovread and LoRDEC hybrid assembly. The most suitable software tool that had the highest corrected Iso-seq data that were supported by RNA-seq evidence was selected.

2.28. Construction of the non-redundant merged transcripts

Transcript resources generated through RNA-seq, Iso-seq and full-length cDNA sequences (Makita et al., 2017) were merged to provide a non-redundant transcript set. Biases that are most inherent to RNA-seq technique are i) the 3'-end bias and ii) over-representation of the long transcripts. This is due to the generic method of most RNA-seq approach that involves sequencing of the fragmented poly(A) RNA sequences. 3'-end bias occurs due to the enrichment of poly(A)-tailed mRNA from total RNA. Furthermore, the 3'-end bias is more pronounced when the mRNA pool is isolated through the usage of a reverse transcriptase to simultaneously isolate mRNA and convert the mRNA into cDNA

(van Dijk et al., 2014). On the other hand, poly(A) purification involving magnetic or cellulose beads coated with oligo-dT molecules, followed by fragmentation of the mRNA and conversion of the fragmented mRNA into cDNA may reduce the 3' end bias (Conesa et al., 2016). In addition to 3'-end bias, length bias may also occur as longer mRNA may have higher cDNA representation , due to the higher amount of fragmented RNA present in the RNA-seq library (Oshlack and Wakefield, 2009). To circumvent the effect of larger size sample from the long transcripts on the downstream analysis, the total number of expressed transcripts will be divided by the transcript's length (Roberts et al., 2011, Patro et al., 2017). This will lead to a normalised transcript-level expression and hence reducing the effect of the bigger sample size of the long transcripts. Therefore, by taking these two measures in the RNA-seq library and transcript profiling analysis, the bias inherent to RNA-seq might be reduced. In addition, the merit of incorporating these independent transcriptome approaches is discussed in Chapter 5, section 5.1. The construction of the non-redundant merged transcripts was performed according to a protocol pipeline as summarised in Figure 2.28.1. The pre-clustering, isoform clustering, gene family reconstruction and isoform collapse (collectively known as merging) were performed by using Coding Genome reconstruction Tool (COGENT) version 3.2 (https://github.com/Magdoll/Cogent). The pre-clustering of the transcripts was carried-out by obtaining the k-mer dictionary and pairwise distances of the transcript sequences. These k-mers and distance profiles were utilised in the gene family reconstruction step, where representation of transcripts for each gene family were generated. Isoform identification was performed in the isoform collapse step, where the merged transcripts were mapped to the transcript representatives. The isoform collapse step was carried out using python scripts

deposited in the cDNA_Cupcake repository (https://github.com/Magdoll/cDNA_Cupcake).

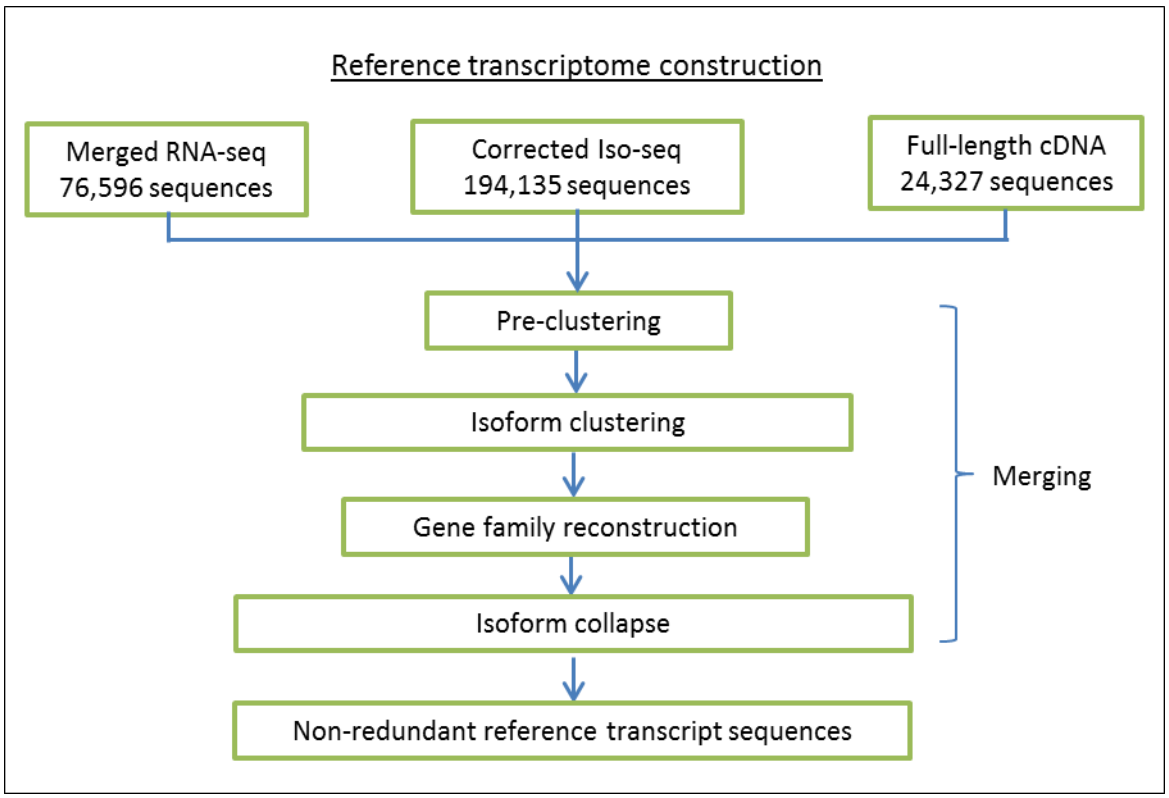


Figure 2.28.1: The outline of processes involved in the generation of the reference transcriptome.

2.29. Evaluation of the merged transcripts

The merged transcripts were evaluated based on three approaches, i) blast analysis; ii) BUSCO evaluation; and iii) analysis of gene family based on key genes in rubber formation. The blast analysis was performed by using the transcriptome as a query against the published annotated transcriptome from *Hevea* (Tang et al, 2016), cassava (https://phytozome.jgi.doe.gov/pz/portal.html) and *Arabidopsis* (https://www.arabidopsis.org/). The blast hit in each case was parsed by keeping the non-redundant hits with e-value ≥ 1 x 10^{-30}, identity percentage ≥ 95

% and query coverage ≥ 50%. The evaluation of essential plant single copy orthologs in the merged transcripts was performed using BUSCO (Simão et al., 2015) using default settings. Prior to viewing of the gene family or isoforms in IGV, the transcriptome was mapped to *Hevea* draft genome (Tang et al, 2016). The mapping was performed using the GMAP aligner (Wu and Watanabe, 2005). The alignment of transcripts to the reference genome was performed using the following settings: f = samse; n = 0; t = 30. The GMAP settings were recommended for mapping of long transcripts to a reference genome (Tseng, 2017). Then, the mapped data was uploaded into Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al., 2013). This allowed for a manual inspection of exon/intron junctions of the targeted gene families. A high coverage and complete transcriptome would exhibit transcripts with a complete coding region and canonical intron/exon junctions.

2.30. Read count generation and differential expression analysis

Prior to differential expression gene analysis, the read count expressed as transcript per million (TPM) for each gene profile was generated. The read count was generated by mapping raw RNA-seq data to the merged, non-redundant transcript reference using Salmon (version 0.8.2) (Patro et al., 2017). The number of reads mapped to the transcriptome was quantified using Salmon's default settings. The read counts generated from the Salmon mapping were summarised into a matrix. The matrix contained information about the count reads at gene level in relation to the RNA-seq data and the matrix was obtained using Tximport R package (Soneson et al., 2016). Afterwards, the differential expression analysis was performed using EdgeR (Robinson et al., 2010), according to a workflow recommended by the software manual.

Following the gene-level summarisation of the read counts, a normalisation factor was calculated using the trimmed median of M (TMM) values method (Robinson and Oshlack, 2010), so that variation within each RNA-seq data could be scaled accordingly. Transcripts with zero or near zero read counts were discarded by applying an expression threshold of TPM > 1. The normalisation of all libraries was re-calculated following the removal of non-expressed and very low-level transcripts. Differential expression analysis was determined by comparing genes from two *Hevea* tree genotypes (RRIM600 and PB235; with PB235 was used as a reference). By using EdgeR algorithm, a gene is considered as differentially expressed if the observed difference showed high dispersion values. The dispersion values for each expressed gene was determined based on the negative binomial model of the count matrix automatically measured by EdgeR. Significant differences in the gene expression between the two *Hevea* tree genotypes were determined based on two thresholds: i) false discovery rate (FDR) < 0.01; and ii) expression value of log fold change (log_2FC) > 1.5.

2.31. Sequence search of *REF* and *SRPP* gene family members

Prior to sequence search analysis, a set of REF and SRPP gene sequences were compiled from the NCBI database, Phytozome version 12.1 and TAIR. To keep to the same labelling convention, the same identifier, REFSRPP was applied to all REF and SRPP gene sequences. The compiled set of REFSRPP sequence is listed in Table 2.31.1. The sequence set was used as query for a TBLASTN analysis (e-value < 1e-10) against the *Hevea* draft genome scaffolds (Tang et al., 2016). To determine corresponding gene models for the TBLASTN hits, the genomic regions flanking the hit sequence (±5000 bp) were lifted from the draft genome sequence. The gene model prediction for the genomic sequence was performed using Augustus (Stanke et al., 2008) and Fgenesh (Salamov and Solovyev, 2000). A manual inspection of the BLAST hits gave a final list of 18 REFSRPP sequences.

The final list of *Hevea* REFSRPP sequences was further confirmed through a reciprocal BLASTN analysis. The reciprocal BLASTN analysis was performed using two datasets; 1) REFSRPP gene models deciphered based on the earlier TBLASTN result; and 2) the genomic sequences from *Hevea* genotypes RRIM928 (Mollison et al., 2014), RRIM600 (Lau et al., 2016) and BPM24 (Pootakham et al., 2017). The reciprocal BLASTN was performed by firstly, using the 18 REFSRPP sequences against the RRIM928, RRIM600 and BPM24 draft genomes. The top hit from the BLASTN results was then used as queries for another BLASTN (e-value < 1e-10) against the Reyan draft genome.

REFSRPP orthologs were identified from *Brassicaceae* (*Arabidopsis, Brassica rapa*), *Euphorbiaceae* (cassava, *Jatropha, Ricinus*), monocots (palm oil, pineapple), and other rubber-producing plants (*Helianthus annuus*, *Lactuca*, milkweed, *Parthenium argentatum* and *Taraxacum*). The sequence search was

performed by using reciprocal BLASTN as mentioned previously) against the genomic resources of the aforementioned plant species.

Subsequently, the predicted coding region of each REFSRPP genes was aligned with known REFSRPP sequences from other species such as cassava, *Arabidopsis*, and published *Hevea* REFSRPP sequences. Manual inspection was performed on the aligned sequences to detect any frame shifts of the coding region or premature stop codons.

Table 2.31.1: A set of non-redundant REFSRPP sequences downloaded from NCBI. Initially, 31 protein sequences were found, and the removal of identical sequences was performed through manual inspection of the alignment of these 31 sequences. The non-redundant sequences were used as queries in the TBLASTN analysis against *Hevea* draft genome (Tang et al, 2016).

Species	**NCBI protein accession number**	**Amino acid length**
Hevea brasiliensis	ALP73392.1	258
	AAP46160.1	117
	AHF95715.1	203
	AAC82355.1	204
	AAR11448.1	175
	AGS13513.1	137
Arabidopsis thaliana	At1g67360	240
	At2g47780	235
	At3g05500	246
Manihot esculenta	Manes.13G012400.1	261
	Manes.12G011500.1	226
	Manes.05G063700.1	236
	Manes.09G170000.1	238
	Manes.08G117800.1	243

2.32. Phylogenetic analysis

A phylogenetic tree was generated by firstly, aligning all translated REFSRPP protein sequences using PRANK (https://www.ebi.ac.uk/Tools/msa/prank/). The aligned sequences were manually inspected to trim regions with missing data. When the alignment showed only 20% of the total aligned sequences, that region was considered as missing data and thus removed from the subsequent analysis. Viewing and editing of the aligned sequences was performed using Jalview version 2 (Waterhouse et al., 2009). The model selection and the subsequent tree construction was performed using TOPALi version 2 (Milne et al., 2009). The phylogenetic tree was viewed using Dendroscope (Huson and Scornavacca, 2012).

2.33. Comparative analysis of the REFSRPP gene family between species

Since the *Hevea* draft genome is not anchored to a genetic linkage map, it is not possible to characterise *Hevea* REFSRPP gene family at the pseudomolecule level. However, some information can be inferred based on the homologous clusters obtained from the REF and SRPP position on the cassava genome. This can be done by comparing neighbouring genes of the orthologous REFSRPP from *Hevea* and cassava. From the phylogenetic tree, it was observed that cassava REFSRPP ortholog (identified as Mes4 in this study) showed the closest phylogenetic relationship to *SRPP9*. Therefore, the homologous clusters were determined based on the matched neighbouring genes surrounding *Mes4* on cassava chromosome and *SRPP9* on *Hevea*

scaffold. The matched neighbouring genes were determined through BLASTN analysis (e-value < 1e-10), using genomic sequences encoding SRPP9 and its flanking genomic sequences (±5000 bp) against cassava chromosome sequences.

2.34. RNA-seq based SNP detection

Bi-allelic SNP calling was performed based on RNA-seq reads mapped on scaffold1222 from *Hevea* draft genome. RNA-seq data from ten *Hevea* genotypes (Table 2.34.1) were used for SNP detection. These RNA-seq reads were mapped to *Hevea* draft genome, according to the pipeline described in sections 2.22 – 2.25. Putative SNPs on scaffold1222 were identified using Freebayes software (Garrison and Marth, 2012). The identified SNPs are supported by at least five alternate bases (for heterozygous SNPs) or 10 bases (for homozygous SNPs). Additionally, SNPs located towards the end of aligned reads (last five bases) were excluded from the SNP calling as this might be due to possible mis-mapping. The SNP polymorphisms identified in all RNA-seq libraries were viewed using Flapjack software tool (Milne et al, 2010).

Table 2.34.1: List of RNA-seq used for SNPs variant calling

***Hevea* tree genotype**	**Tissue**	**NCBI Accession no.**
Reyan	root	SRR3136156
	bark	SRR3136158
	leaf	SRR3136159
	latex	SRR3136162
	female flower	SRR3136165
	male flower	SRR3136166
	seed	SRR3136168
RRIM600	petiole	DRR069093
	leaf	DRR069092
	latex	DRR069091
	bark	DRR069090
PB 235	latex	Own library*
PR107	bark	SRR2147074
	bark	SRR2147073
	bark	SRR2147072
	leaf	SRR3240371
	latex	SRR2001614
	bark	SRR1611790
	bark	SRR1588170
	latex	SRR1533844
RRIM 928	leaf	SRR854522
	bark	SRR854520
	latex	SRR854521
BPM 24	latex	SRR1649353
RRIC 110	latex	SRR1649349
RRII 105	latex	SRR1649350
RRIM 901	latex	DRR069094
BPM 24	leaf	Not applicable**

*: RNA-seq dataset for PB 235 were obtained from the cDNA libraries of total RNA from latex samples collected in the present study.
**: RNA-seq from BPM 24 was obtained from http://www4a.biotec.or.th/rubber/

2.35. Genomic-based marker design and KASP genotyping

The SNP markers on the *REFSRPP*-rich genomic scaffold were detected based on the alignment of scaffold1222 to the *Hevea* draft genome sequence assembled from genotypes Reyan, RRIM928 and RRIM600. Subsequently, primers targeting 201-bp amplicons containing these SNP markers were

designed, so that the primers could be used to genotype *Hevea* trees through the Kompetitive allele-specific PCR (KASP) assay (LGC Genomics, UK) protocol. The design and final selection of SNP markers to be included in KASP assay optimisation was based on the recommendations by Dr Kelly Houston, from the Cell Molecular Sciences group, the James Hutton Institute. The primer sequences, location and allelic information are summarised in Figure 2.35.1.

The SNP markers were used for genotyping of 51 *Hevea* tree genotypes using KASP assays. These DNAs are part of an archive of rubber genomic DNA resource created by Dr Keng-See Chow (Malaysian Rubber Board) since 2011. Leaves of 50 genotypes were collected mainly from clonal source bushes planted in a germplasm block (Field 68) of the Malaysian Rubber Board collection. In addition, leaf sample from Reayn 77-33-97 was contributed by Dr Chaorong Tang, of Chinese Academy of Tropical Agricultural Sciences (CATAS), Haikou, China. Leaf genomic DNA was prepared using the NucleoSpin Plant II Kit (Machery Nagel, USA). The DNA samples were transported from the Malaysian Rubber Board by international courier to the James Hutton Institute for the KASP assay. The DNA templates were diluted to a final concentration of 10 ng/µl and a minimal 40 ng of the diluted DNA was used in the KASP assay. The KASP assay was performed using KASP reaction master mix (version 3 chemistry, LGC Genomics, UK) as per the manufacturer's instruction, in the StepOnePlus™ Real-Time PCR System (Applied Biosystems, US). Genotype caling from the KASP assay results was performed using StepOne Software v2.1 (Applied Biosystems, UK).

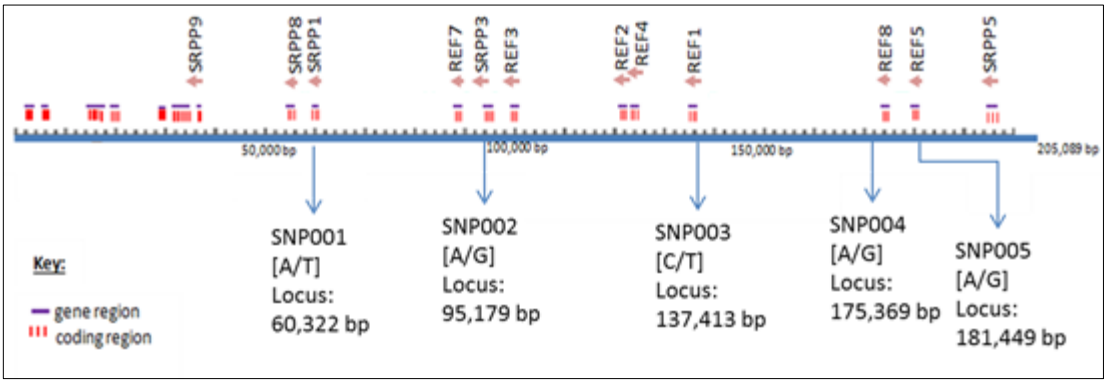


Figure 2.35.1: A schematic diagram of scaffold1222 (length = 205,089 bp) assembled from Reyan 7-33-97 *Hevea* tree clone genome (Tang et al, 2016). The selected SNPs (SNP001-005; its locations are as illustrated) were identified from the overlapped regions of Reyan 7-33-97-RRIM600-RRM928 DNA sequences. The corresponding primers for the selected SNPs were designed so that 100 bp region would flank the targeted variant.

Chapter 3

Carotenoids identification and quantification by HPLC-DAD-MRM from the latex of *Hevea brasiliensis*

3.1. Brief Introduction

3.1.1. Carotenoid formation in rubber

Carotenoids are important secondary metabolites, synthesized by all photosynthetic organisms (Young, 1991, Alagoz et al., 2018). In higher plants, carotenoids are stored in underground organs such as tubers or roots and all type of plastids such as chromoplasts and chloroplasts (Li and Yuan, 2013, Howitt and Pogson, 2006, Morris et al., 2004, Nisar et al., 2015). In *Hevea* latex, carotenoids are found to be accumulated in the Frey-Wyssling (FW) particles (Gomez and Samsidar, 1989). In addition, differential carotenoid accumulation is observed among different type of plastids (Ruiz-Sola and Rodríguez-Concepción, 2012). Indeed, in *Hevea* latex, its yellow hue is mainly dependent on the number and size of the FW particles and accumulated carotenoid pigment inside the particles (Gomez and Samsidar, 1989; Cockbain and Southorn, 1962)

Although the mechanism of carotenoid synthesis has never been described in *Hevea* latex, its formation in other plant species is well characterised. Similar to the formation of rubber (described in Chapter 1, section 1.3), carotenoid synthesis starts from the formation of isopentenyl diphosphate (IPP). However, in contrast to rubber that predominantly uses farnesyl diphosphate (FPP) in their synthesis route, carotenoid biosynthetic steps utilise geranylgeranyl diphosphate (GGPP) (Hirschberg, 2001). There are two distinct stages in the carotenoid biosynthesis pathway, i) lycopene synthesis and ii) the xanthophyll cycle (Figure 3.1.1.1). The lycopene synthesis commences with the formation of phytoene from GGPP, mediated by phytoene synthase (PSY). PSY is reported to be the rate-limiting enzyme in the

carotenoid biosynthesis pathway (Hirschberg, 2001, Fraser and Bramley, 2004, Cazzonelli and Pogson, 2010). The five enzyme-mediated steps end with the production of lycopene. The first step of the xanthophyll cycle begins with the lycopene molecules being added with two different cyclic end groups i.e. beta or epsilon. The cyclization of lycopene into either β-carotene or α-carotene is mediated by two lycopene cyclases, β-cyclase (βLCY) and ε-cyclase (εLCY) (Cunningham and Gantt, 2001). The βLCY enzyme mediates the formation of β-carotene and its derivatives. The εLCY is involved in the synthesis of α-carotene and its derivatives.

In *Hevea* laticifers, IPP synthesis takes place in the cytosol via the cytosolic MVA pathway and in plastids via the MEP pathway (Chow et al., 2012). IPP cross-talk between plastids and cytosol in latex has been reviewed in Chapter 1, section 1.3. Apart from the hypothesis that MEP-derived IPP might be used in rubber biosynthesis in mature *Hevea* trees, no further studies have been reported on the utilisation of plastidic and cytosolic IPP by different types of latex isoprenoids. Considering the lack of knowledge on the specific pathway which provides IPP for rubber and non-rubber isoprenoid formation, understanding the regulation that underlies latex isoprenoid synthesis is of considerable interest.

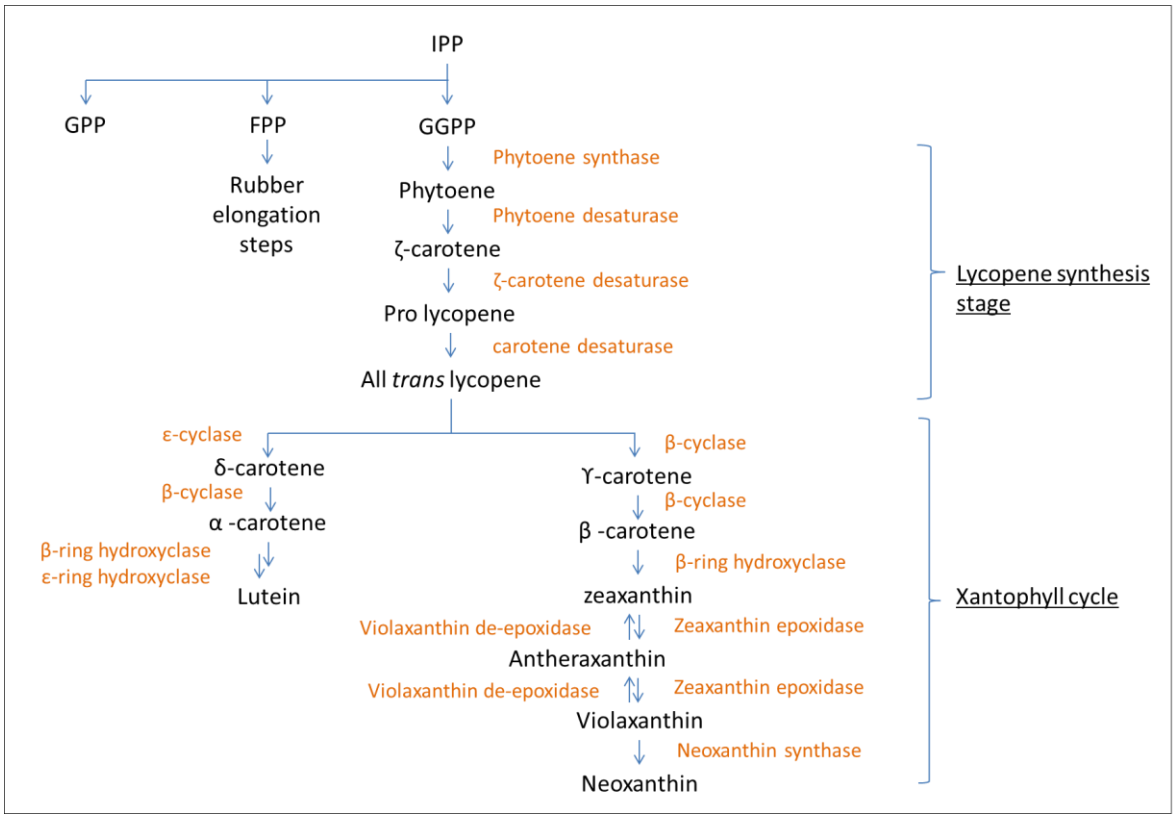


Figure 3.1.1.1: Plant carotenoid biosynthesis. The synthesis of carotenoid starts with the IPP generation step, leading to the formation of the lutein and neoxanthin.

3.1.2. Carotenoids affecting the aesthetic property of processed latex

Although the light-colour of processed rubber does not impact its physical properties, light-colour processed rubber is of a premium grade (Dulngali and Ong, 1980, Boonyang and Sakdapipanich, 2014). The darkening of rubber products is believed to be caused by either pigment oxidation or an enzymatic reaction mediated by polyphenol oxidase (Boonyang and Sakdapipanich, 2014). Carotenoid have been identified as one of the pigments that cause the darkening of processed rubbers (Eaton and Fullerton, 1929, Sakdapipanich, 2006). To produce light-colour processed rubber, a combination of the following procedures is followed: 1) blending the clonal latex samples so that they result in intermediate colour values; 2) usage of minimal ammonia and addition of antioxidant agent such as sodium bisulphite; and 3) rinsing of the processed rubber with a large amount of water, to prevent oxidative darkening (Dulngali and Ong, 1980, Boonyang and Sakdapipanich, 2014). These procedures will incur additional rubber processing cost and involve the generation of a large volume of water effluent (Devaraj et al., 2013). Information on the regulation of carotenoid synthesis and its accumulation in *Hevea* latex plastids is limited. Such information is important for understanding the biochemical and molecular basis of clonal variations in carotenoid content. Subsequently, this may be applied in developing sustainable breeding and supply chain strategies to provide processors with light-colour dry rubber.

3.1.3. Latex coagulation

Latex collected from rubber plantations will naturally pre-coagulate within 5- 6 hours after tapping (John, 1976, Abraham and Hashim, 1983, Salomez et

al., 2014). The non-reversible coagulation causes rubber particles in latex to separate from the latex serum and form networks of tightly-bound rubber coagulate (Woo, 1973). The coagulated rubber also contains a variable amount of entrapped serum (Bauer et al., 2014). Latex coagulation is due to changes in rubber particle membrane stability once it is expelled from the tree upon tapping. This is because lutoids in tapped latex are similarly destabilised and ruptured lutoids result in a drop of pH. The latex will remain fluid if the surface of the latex organelles are negatively-charged (Southorn and Yip, 1968). In contrast, if these negative charges are neutralised, the particle membrane will destabilise hence promoting the aggregation or coagulation of the rubber particles. The charge neutralisation occurs when latex pH drops, which then ruptures the particle membrane and causes rubber particles to fuse into the irreversible rubber coagulates (Southorn and Yip, 1968). To circumvent latex coagulation, the most common anti-coagulant used is ammonia (NH_3) solution (Cook, 1960, Bauer et al., 2014). The addition of ammonia increases the alkalinity of latex serum hence maximises latex stability (John, 1976).

It has been observed that the physical and chemical properties of the preserved latex contrast significantly with those of the freshly tapped latex. For example, the addition of ammonia was reported to negatively affect the integrity of the rubber organelles and protein content (Subramaniam, 1980, Yeang et al., 2002, Kekwick et al., 1996). The addition of ammonia is implicated in increasing a higher anionic strength of latex and hence, it affects membrane stability of latex particles (Woo, 1973, Ho, 1989). This is evidenced when the ammoniated latex is fractionated through high-speed centrifugation, the bottom fraction is totally eliminated (Yeang et al., 2002). Despite accumulating evidence of the effect of ammonia addition on the physio-chemical properties of *Hevea* latex, no

such study has been done on the effect on latex metabolite content. Therefore, it is advisable to directly use fresh latex without involving the usage of ammonia preservation in the analysis of latex metabolites.

3.1.4. Aims

In an effort to identify and measure carotenoid composition in *Hevea* latex, samples were collected from RRIM600 and PB235 rubber tree genotypes. As described in Chapter 1, section 1.3, PB235 *Hevea* genotype produces latex which is distinctly yellow and the RRIM600 *Hevea* genotype generates creamy-white latex. The different latex colour indicates differential accumulation of carotenoid in *Hevea* latex. As fresh latex will coagulate within a few hours after harvesting, it is necessary to work with fresh latex and collect it under controlled (cold) conditions. Since the addition of ammonia into the harvested latex may compromise its metabolite composition, transporting fresh, fluid latex from over a long distance is not an option. Carotenoid extraction from latex has to be performed on fresh latex, *in situ* and with minimum delay after collection. Subsequently, to show differential accumulation in the collected latex, total carotenoids were estimated by UV-visible absorbance (at 450 nm) and the identification and quantification of the individual components were completed using HPLC-MS/MS.

3.2. Results

3.2.1. Optimisation of carotenoid extraction from *Hevea* latex

3.2.1.1. Selection of the best working solvent

To date, little information is available regarding carotenoid profiling from the latex of *Hevea brasiliensis.* On the other hand, the characterisation of total lipids from *Hevea* latex has been well documented (Ho et al., 1975a, Hasma, 1991, Liengprayoon et al., 2008, Bonfils et al., 2007). The most common method to extract *Hevea* latex lipid is to use a Soxhlet extraction apparatus with chloroform-methanol, based on a procedure described by Folch et al. (1956). However, Bonfils et al. (2007) found that this procedure often gave inconsistent extraction yield. Furthermore, the relatively high temperature used in the Soxhlet apparatus operation will degrade sensitive compounds such as carotenoids (Boon et al., 2010). At the same time, chloroform used in the protocol could dissolve rubber chains (located inside rubber particles) and the dissolved rubber would persist in the final extraction product (Hawkins, 1984). Therefore, a selection of solvent that can efficiently extract carotenoid from latex has to be carried out.

In this assessment, carotenoid extraction from latex samples was performed using three different solvents, namely acetone, petroleum ether and tetrahydrofuran. It was observed that latex samples in acetone formed rubber coagulate which sedimented as a pellet phase (after a centrifugation separation at 12,000 g for 15 minutes) (Figure 3.2.1.1.1). On the other hand, the rubber particles were coagulated when added into petroleum ether. Then, after a few hours, some of the coagulated rubber were found to be solubilised into the solvent phase. When latex samples were added into tetrahydrofuran, the rubber content

solubilised completely into the solvent phase and no rubber coagulate was detected.

Due to the solubilised rubber, any assessment of the UV-visible spectral characteristics of the solvent phase leads to incorrect estimation of total carotenoids in the extraction product. A good solvent for carotenoid extraction therefore must be able to remove most of the carotenoids from latex whilst not dissolving the rubber chains. Based on the feasibility of recovering the solvent phase from rubber coagulates, acetone was selected as a working extraction solvent for the subsequent step of obtaining carotenoid from latex.

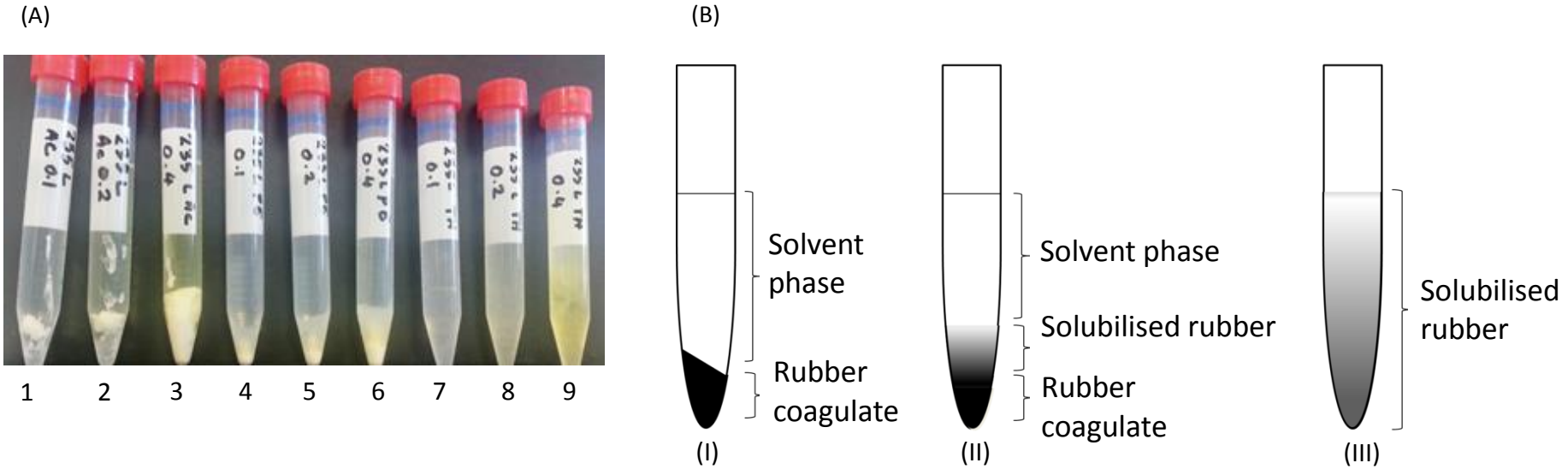


Figure 3.2.1.1.1: The separation of solvent and rubber phases during carotenoid extraction from latex samples. Tinge of yellow in the solvent fractions indicates that carotenoid had diffused out from latex matrix.

(A) indicates the extraction of carotenoid using acetone (tubes 1-3), petroleum ether (tubes 4-6) and tetrahydrofuran (tibes 7-9). For each extraction, three different latex volumes were used in 100% solvent; 100 µl (tubes 1,4 and 7); 200 µl (tubes 2,5 and 8); and 400 µl (tubes 3,6 and 9). (B) shows the schematic diagram of the rubber and solvent phases after being incubated for 6 hours.

(B):(I) shows the layers of rubber and solvent layers clearly separated in acetone extraction. (II) shows the extraction using petroleum ether. Three different fractions were observed for latex in 100 % petroleum ether. The fractions consisted of solvent phase, solubilised rubber phase and rubber coagulum pellet. (III) shows the extraction using tetrahydrofuran. Rubber particles were observed to be fully dissolved in 100% tetrahydrofuran.

3.2.1.2. Extraction recovery

The purpose of this experiment was to determine the best method for extracting carotenoids quantitatively from latex samples. This would enable comparison of latex carotenoid content between different *Hevea* genotypes. In this study, latex samples from PB235 and RRIM600 rubber trees had to be transported from the harvest site in Malaysia. Initially, the feasibility of transporting frozen fresh latex was investigated as storing at low temperature might help to conserve latex metabolites at the time of its harvesting. However, the frozen latex will become coagulated after thawing process and the aggregated rubber particles in the coagulated latex may entrap non-rubber compounds within the latex matrix. This may hinder full recovery of non-rubber compounds from the coagulated latex.

Therefore, the extraction recovery assessment was performed on the coagulated and fluid non-ammoniated latex (fresh latex). The extraction recovery study was achieved by comparing the extracted amount of a carotenoid control spiked into coagulated and fluid latex. The carotenoid compound used for spiking was *trans*-β-Apo-8′-carotenal. The compound was isolated as a minor carotenoid compound from citrus fruits (Stewart and Wheaton, 1973, Matsumoto et al., 2007) and therefore, is not expected to exist naturally in latex. The compound has also been used as an internal standard in other plant carotenoid profiling studies (Lashbrooke et al., 2010, Wang et al., 2014).

The extraction starts when the latex sample is added to acetone and stirred for 16 hours. The rubber-acetone mixture was separated into rubber pellet and yellow solvent phase after a centrifugation step. Examination of the latex pellet after 16 hours revealed that only traces of yellow could be observed on the pellet surface. Hence, it was assumed that the removal of the majority of

carotenoids was achieved from the latex matrix. However, to ensure most of the carotenoids were fully extracted, the extraction was repeated twice (the extraction time was reduced to 3 hours, instead of 16 hours), so that three batches of yellow solvent fractions were obtained. These fractions were separated by chromatography based on the reverse-phase HPLC protocol described by Fraser et al (2000).

The percentage of the recovered spiked carotenoid was calculated by comparing its HPLC peak area to that of its calibration curve. The calibration curve indicated linearity over a range of ten known concentrations (between 0.5 part per million (PPM) to 100 PPM). The calculated r^2 value was above 0.997. The recovery percentage of the spiked *trans*-β-Apo-8′-carotenal from sequential reaction of the latex samples is summarised in Table 3.2.1.2.1. The first, second and third sequential extractions from the fluid latex gave 93.17, 3.11 and 2.25 % recovery of the spiked carotenoid. On the other hand, a lower recovery (46.37, 0.27 and 0.05%) was obtained from the coagulated latex. Therefore, it was established that to get an accurate estimation of latex carotenoids, it is highly recommended to use fluid latex as a starting material as it showed a significantly higher extraction yield.

Table 3.2.1.2.1: The minimal recovery percentage (± standard error) of spiked *trans*-β-Apo-8′-carotenal extracted from latex samples. The calculations were obtained from 6 technical replicates of the extractions.

Sequential extraction time	**Mean recovery percentage (%)**	
	Liquid latex	Coagulated latex
A: Latex sample stirred in acetone for 16 hours	93.2 ± 7.66	46.4 ± 1.09
B: Pellet from A, stirred in acetone for 3 hours	3.1 ± 0.37	0.3 ± 0.02
C: Pellet from B, stirred in acetone for 3 hours	2.3 ± 1.73	<0.1

3.2.2. Estimation of total carotenoid content from RRM600 and PB235 latex samples

The purpose of this experiment was to calculate total carotenoid contents from the *Hevea* latex. This would give an overview of carotenoid accumulation in the *Hevea* latex. Latex samples collected from the two rubber tree genotypes, RRIM600 and PB235 used as a basis for differential carotenoid content in this study. During the extraction process with acetone, the carotenoid diffused out from latex matrix and turned the solvent phase into a clear, yellow phase. This is due to the conjugated double bonds within the polyene chain of carotenoid compound (Weedon and Moss, 1995). In addition, such molecular structure allows absorption of visible light. Therefore, carotenoids can be measured through the light absorption spectrum. The estimation of carotenoids extracted from the latex of RRIM600 and PB235 were determined through spectrophotometer absorbance, as described in Chapter 2 (section 2.7).

The evaluation of total carotenoids from the fluid latex of RRIM600 and PB235 is summarised in Figure 3.2.2.1. Total carotenoids in the latex of RRIM600 and PB235 were estimated to be at 1.16 and 5.10 µg/g, respectively.

The higher amount of total carotenoids in PB235 latex further confirmed the assumption that higher accumulation of carotenoid is associated with yellow latex.

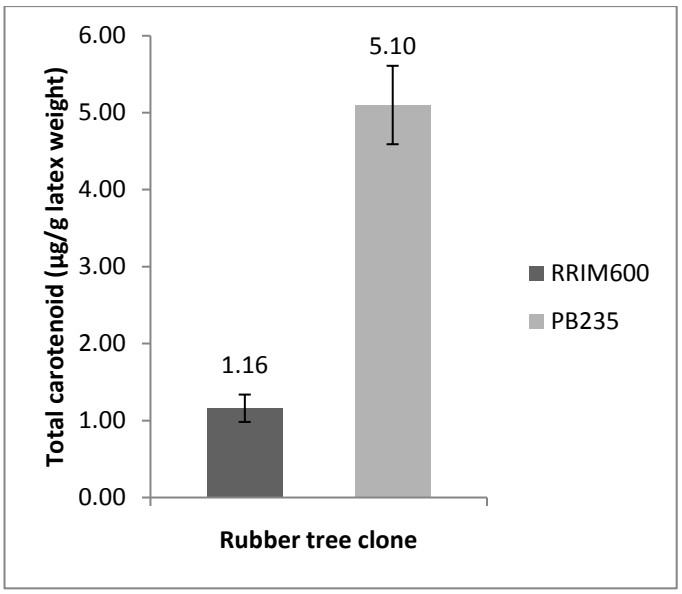


Figure 3.2.2.1.: Total carotenoid contents (*p*-value < 0.05), estimated by spectrophotometric analysis for fluid latex collected from RRIM600 (*n* = 6) and PB235 (*n* = 4). Error bars represent the standard error values. P-values which indicate the significance of the difference of carotenoid content between PB235 and RRIM600 were calculated using unpaired Student's *t*-test.

3.2.3. Dry rubber content of yellow and white latex samples

The objective of this experiment was to estimate the rubber content of PB235 and RRIM600 latex samples. By estimating the rubber content, it would facilitate the comparison of polyisoprenoids content between PB235 and RRIM600 latex samples. The dry rubber content (DRC) measurement was executed as described in section 2.13 in Chapter 2.

Although it is a composite calculation, it is assumed that DRC value represents the hydrocarbon content of latex as rubber is the major component of *Hevea* (D'Auzac and Jacob, 1989). Additionally, in the biochemical studies of enzymes involved in rubber biosynthetic pathway, DRC value has been routinely used to reflect rubber content changes in *Hevea* latex samples (Archer et al., 1969, McMullen and McSweeney, 1966, Suwanmanee et al., 2013) and from other rubber-producing plants (Schmidt et al., 2010b). Higher DRC values (>41%) have been linked to the higher capacity of a rubber tree producing natural latex (Eng et al, 2001). The DRC measurement generated from the yellow and white latex samples used in this study is summarised in Figure 3.2.3.1. It indicates that rubber content of RRIM600 latex is higher (average DRC = 50.37%) than that of PB235 (average DRC = 46.11%).

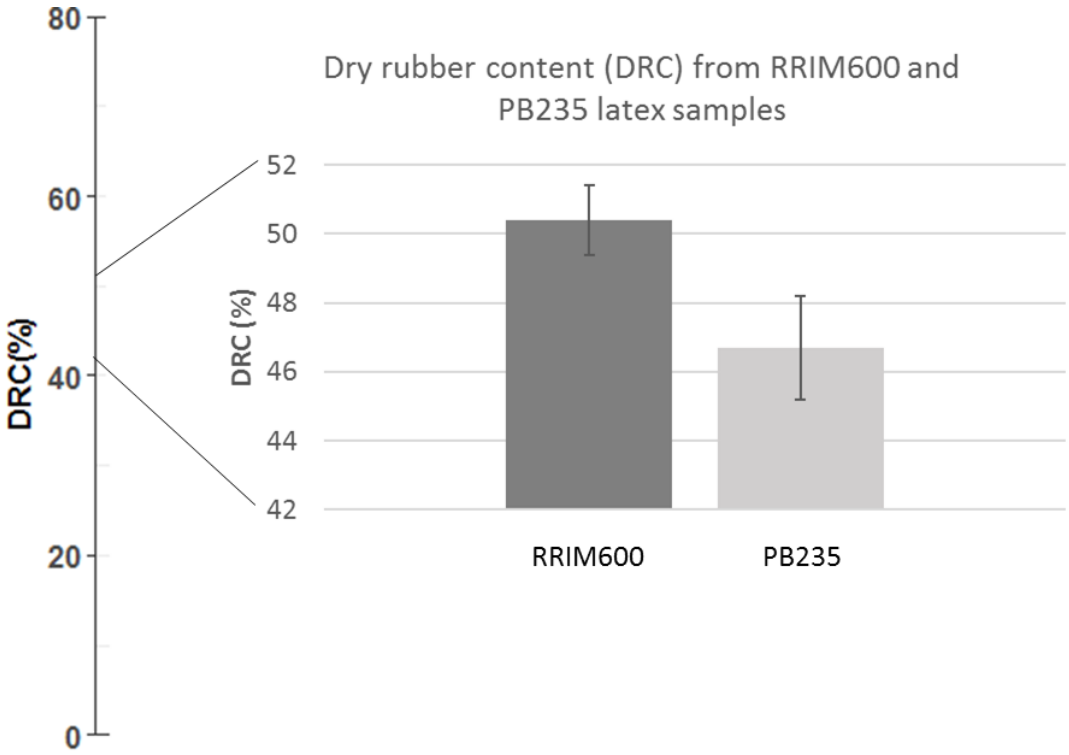


Figure 3.2.3.1: Dry rubber content measurement for PB 235 (n=4) and RRIM 600 (n =6) latex samples ($p < 0.005$). Overall, the DRC from the samples fall above the high DRC values (>41%). P-values which indicates the significance of the difference of dry rubber content between PB235 and RRIM600 were calculated using unpaired Student's *t*-test.

3.2.4. The separation and identification of carotenoids from the latex of *Hevea brasiliensis*

The analysis was performed to characterise the carotenoid constituents in the *Hevea* latex. Based on the past studies, it had been observed that carotenoid compounds were extracted with other non-rubber constituents, which collectively known as the latex total lipid (Ho et al., 1975a, Hasma and Subramaniam, 1986). Additionally, Sakdapipanich (2006) identified β-carotene in *Hevea* latex and confirmed its structure by Fourier transform infra-red spectroscopy (FTIR). Apart from these reports, no information pertaining to the usage of a suitable analytical method for latex carotenoid profiling is available.

In this study, latex carotenoids composition was characterised by a combination of high-performance liquid chromatography (HPLC), coupled to a diode array detector (DAD), based on an analytical procedure recorded by Fraser et al. (2000). The method is described in detail in section 2.8 in Chapter 2. The HPLC-DAD run was performed by injecting individual sample of the latex extraction (both saponified and non-saponified) from the RRIM600 and PB235 latex samples and the corresponding carotenoid authentic standards. The saponification process involves alkali hydrolysis of esterified carotenoids and removes contaminating substances such as neutral lipids from latex extracts (Schiedt and Liaan-Jensen, 1995). A complete saponification will generate free carotenoids and hence, facilitates the HPLC peak identification. The saponification of latex extracts was performed as described in section 2.6 in Chapter 2.

A typical HPLC chromatogram profile of the *Hevea* latex is shown in Figure 3.2.4.1. Qualitatively, the chromatographic pattern was similar for both

RRIM600 and PB235 latex samples. No significant difference was detected between the non-saponified and the saponified extraction products; other than a distinct decrease of peak 6 in the saponified extracts (Figure 3.2.4.2). This indicated peak 6 contains esterified carotenoids. On the other hand, it was observed that the signal intensity of the HPLC chromatogram for peaks in RRIM600 was consistently lower than the signals produced by compounds from PB235 latex. This correlated to the previous observation that total carotenoid contents in the latex of PB235 are higher compared to RRIM600.

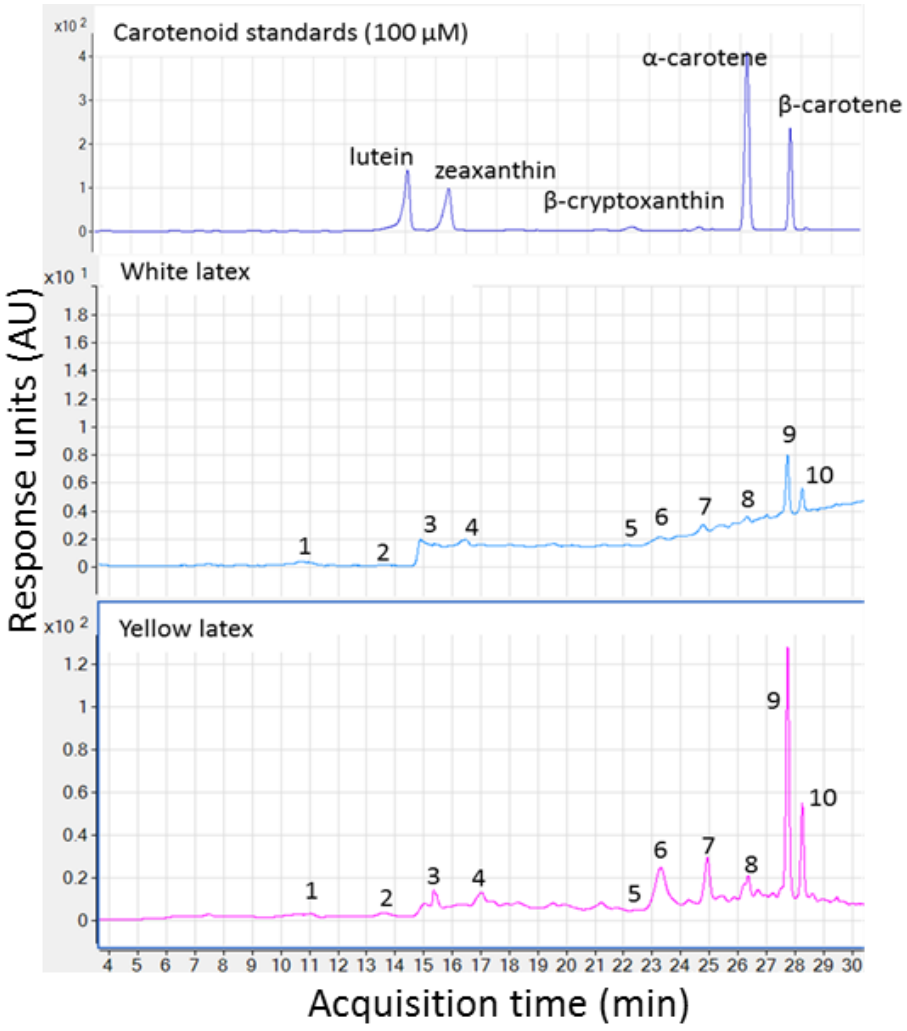


Figure 3.2.4.1: A typical chromatographic pattern viewed at 451nm, generated from the reverse-phase HPLC-DAD of latex samples; RRIM600 (white latex) and PB235 (yellow latex). All compounds were eluted within 30 minutes during the HPLC run. 1: Neoxanthin; 2: unknown; 3: lutein; 4: zeaxanthin; 5: β-cryptoxanthin; 6: unknown; 7: unknown; 8: α-carotene; 9: β-carotene; 10: unknown.

The acquired data from the HPLC-DAD analysis contained the following information: 1) the retention times for each HPLC peak; 2) HPLC chromatogram elution order; and 3) spectral characteristics of the eluents. By comparing the acquired data from the latex samples to the evidence of the corresponding authentic standards, carotenoid identity was assigned to each HPLC peak. The spectral characteristics for carotenoids are defined by their 3-peak-absorption curves and by comparing them to the published carotenoid's absorption curves, the identity can be inferred. The 3-peak-absorption curves for peaks generated from the HPLC separation of latex extracts is shown in Figure 3.2.4.2 and the spectral characteristics for latex carotenoids are summarised in Table 3.2.4.1. Six carotenoids namely neoxanthin, lutein, zeaxanthin, β-cryptoxanthin, α-carotene and β-carotene were assigned to the HPLC peaks, based on the supporting data as described above. However, due to the low level of β-cryptoxanthin, its corresponding UV-visible spectral characteristic could not be accurately determined. Hence its identity was only inferred based on comparing retention times of the compound to its corresponding standard.

Additionally, the annotation of the tandem mass spectrometry (MS/MS) lent further support to the identification of carotenoid compounds. In the MS/MS run, two types of information were generated, 1) full MS that infers the mass of the molecular ion and 2) MS/MS that infers the mass of the ion fragments. The *m/z* values of both molecular ion and ion fragments can be used for inferring the molecular structure of compounds of interest.

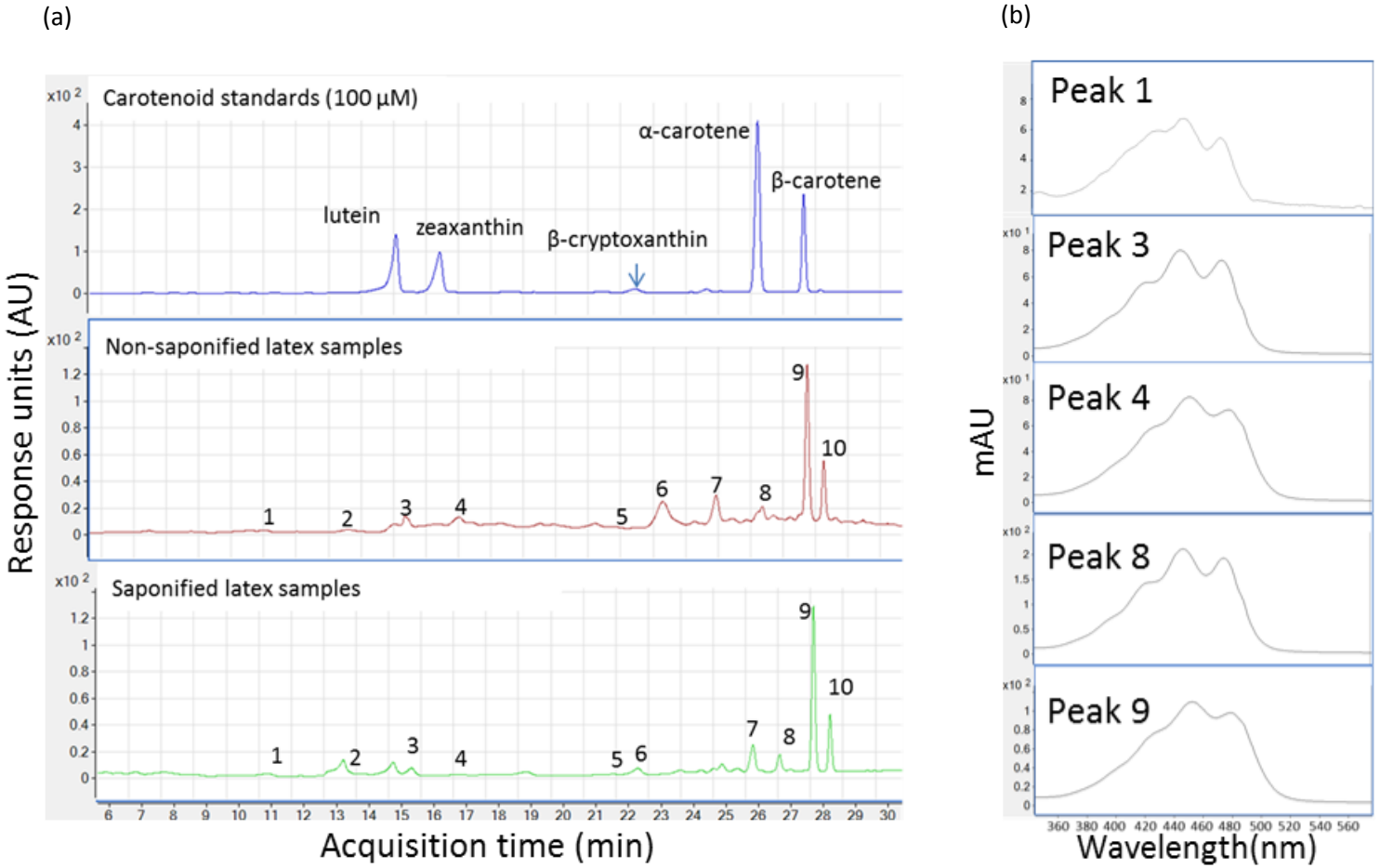


Figure 3.2.4.2: (a) A typical chromatographic pattern generated from the reverse-phase HPLC of non-saponified and saponified latex sample. Putative carotenoid peaks were labelled from 1 to 10 (1: Neoxanthin; 2: unknown; 3: lutein; 4: zeaxanthin; 5: β-cryptoxanthin; 6: unknown; 7: unknown; 8: α-carotene; 9: β-carotene; 10: unknown.
(b) Derived spectral characteristics from the non-saponified latex samples for the identified carotenoids.

Table 3.2.4.1: The spectral characteristics and the MS profiles of HPLC chromatographic peaks obtained from RRIM600 and PB235 latex extraction products.

Peak	**Compounds**	**Experimental UV spectral (at 451 nm)**	**Published UV spectral (Fraser et al, 2000)**	**Molecular species**	**Full scan (*m/z*)**	**MS/MS fragment ions (*m/z*)**
1	Neoxanthin	415.0, 434.0, 465.0	415.0, 440.3, 466.9	$[M]^{+\bullet}$	601	167,119,105
3	Lutein	420.0, 440.0, 473.0	421.0, 443.9, 477.8	$[M]^{+\bullet}$	568	476,119,105
4	Zeaxanthin	426.0,448.0,475.0	425.0, 451.2, 477.8	$[M]^{+\bullet}$	568	476,119,105
5	β-cryptoxanthin	**not detected	430.0, 451.2, 477.8	$[M]^{+\bullet}$	552	**not detected
8	α-carotene	420.0, 443.0, 470.0	423.0, 446.3, 474.1	$[M]^{+\bullet}$	536	444, 119, 105
9	β-carotene	-, 452.0, 478.0	-, 453.0, 478.0	$[M]^{+\bullet}$	536	444, 119, 105
2	unknown	410.0, 435.0, 465.0	not applicable			
6	unknown	420.0, 445.0, 475.0	not applicable			

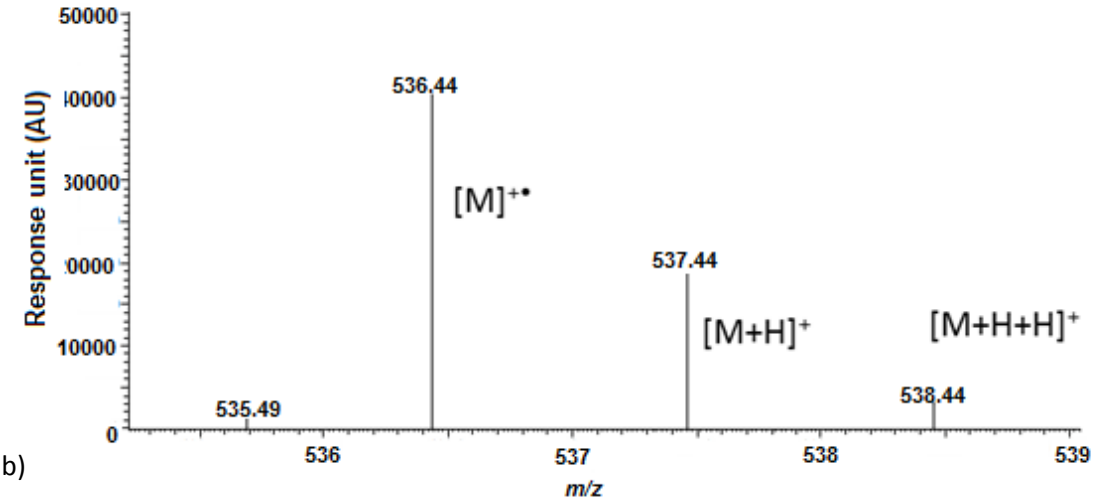
7	unknown	420.0, 435.0, 464.0	not applicable
10	unknown	426.0, 440.0, 470.0	not applicable

**The compound was too low; the compound identification was based on *m/z* value and retention time of its corresponding standard

The molecular ion mass for carotenoids identified from the latex samples is detailed in Table 3.2.4.1. Although the ionisation process was performed in positive mode for latex carotenoids, the molecular ions were observed as radical ions, instead of forming protonated ions. The occurrence of radical ions for carotenoid compounds when ionised in positive mode has been reported previously (Fraser et al., 2008, Rivera et al., 2014). Due to the settings used for MS/MS operation in this study (which only scanned the three most abundant ions) only the MS/MS profile for β-carotene was obtained (Figure 3.2.4.3). Other carotenoid ions were too low in abundance to be detected by the MS scanner. From the fragmentation pattern, the most intense fragment ion for β-carotene was observed at *m/z* 444. This indicates the loss of toluene structure from β-carotene hydrocarbon backbone (*m/z* $[M-92]^{+\bullet}$) (Lacker et al., 1999). The proposed mechanism of the toluene breakage is outlined in Figure 3.2.4.3. The breakage of toluene during carotenoid fragmentation was generally observed in most carotenoid compounds (Rivera et al., 2014).

This is the first analysis using HPLC-DAD-MS/MS that has ever been reported to separate and identify carotenoid compositions from *Hevea* latex. It provides an overview of the typical carotenoid content of *Hevea brasiliensis* latex. Based on the separation and the identification of carotenoids extracted from the RRIM600 latex (n=6) and PB235 latex (n=4), it was observed that lutein, zeaxanthin, α-carotene and β-carotene were consistently amongst the major compounds present in both RRIM600 and PB235 latex samples. Therefore, a targeted quantification using HPLC was subsequently performed for these four compounds.

(a)



(b)

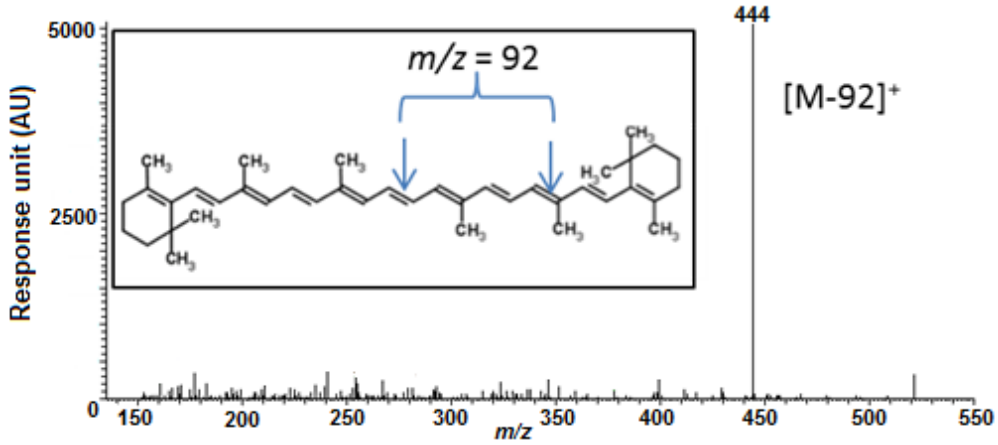


Figure 3.2.4.3: Mass spectra for β-carotene found in *Hevea* latex.
(a) Full scan for MS in positive mode, generating the most abundance molecular ion at *m/z* = 536.0 ($[M+H]^{\bullet +}$). It also produced smaller signals for protonated ions at *m/z*= 537.0 ($[M+H]^{+}$) and 538.0 ($[M+H+H]^{+}$).
(b) Fragmentation of β-carotene through MS/MS producing fragment ion at *m/z* = 444.0. Inset: probable fragmentation points (as indicated by the arrows) for β-carotene during the ionisation process.

3.2.5. Quantification of major carotenoids

This experiment was executed with the aim of accurately quantifying the major carotenoids from PB235 and RRIM600 latex samples. This enables a comparison of carotenoid compositions both qualitatively and quantitatively between yellow and white latex samples. In this study, following the separation and identification of carotenoid compositions from the RRIM600 and PB235 latex samples, it was found that β-carotene, α-carotene, lutein and zeaxanthin were consistently present in latex samples. Consequently, a targeted quantification of the four compounds was carried-out by high performance liquid chromatography-diode array detector-multiple reaction monitoring (HPLC-DAD-MRM).

Most carotenoid pigments contain a chromophore structure, which is defined as double bonds between carbon-carbon in the carotenoid backbone and the carbons are arranged in a conjugated system. This is what gives carotenoids its colour and spectral characteristics (Britton, 1995). During the HPLC-DAD analysis, the chromophore structure from eluted carotenoids will absorb light and this can be detected through diode array detector, which then can be used for the quantification purposes. Subsequently, the separated carotenoids will enter the triple quadrupole for ion fragmentation during the MRM execution. Rather than performing a full MS scan for the analytes, specific *m/z* settings (often called transition) are used to detect the compounds of interest. By just focussing on a set of targeted transitions, the efficiency and sensitivity of the HPLC-DAD-MRM can be increased. This is amenable to detecting both low-abundance and high-level metabolites. Previously, a full mass spectra scan of latex extracts only produced an MS/MS profile for β-carotene. Other carotenoids, namely lutein, zeaxanthin and α-carotene could

not be detected by the detector, due to their low abundance. Therefore, by employing HPLC-DAD-MRM, it facilitates the identification and quantification of varied amount of carotenoids in the latex of *Hevea brasiliensis.* The HPLC-DAD-MRM used for identification and quantification of β-carotene, α-carotene, lutein and zeaxanthin was carried out as described in sections 2.8 and 2.10, Chapter 2.

To ensure the analytical protocol is highly sensitive to detect both the highest and lowest carotenoids from latex extracts, the transitions used for the MRM settings have to be optimised. Following the MRM optimisation, all targeted carotenoids exhibited two common transitions, which contain abundant ions at *m/z*= 105 and *m/z*=119. This is due to the similar basic structure shared by carotenoids, which is represented by the C_{40} carbon backbone chain (Britton, 1995). Similarly, all targeted carotenoids produced a high level of $[M-92]^+$ product ions, corresponding to the loss of toluene from the polyene chain in the carotenoid structure, which had been described earlier in the MS/MS result of β-carotene. Based on the intensity level, these three product ions (*m/z*= 105, *m/z*=119 and $[M-92]^+$) were chosen as the MRM transitions for the targeted carotenoids monitoring. Essentially, α-carotene and β-carotene detection were executed by monitoring precursor ion *m/z* = 536 and the product ions of *m/z*=105, *m/z*=119 and *m/z*=444. The identification of lutein and zeaxanthin was supported by detecting precursor ions *m/z* = 536 and the product ions of *m/z*=105, *m/z*=119 and *m/z*=476. Identical transitions were used to detect α-carotene and β-carotene as both compounds are isomers and hence share similar molecular structure. However, it was observed that α-carotene and β-carotene exhibited differential retention times, and therefore, the transitions would still give meaningful evidence for compound identification. Likewise,

lutein and zeaxanthin are isomers and although identical transitions were used, identification could still be carried-out as both compounds were eluted more than one minute apart from each other. The identified HPLC-DAD-MRM peaks of lutein, zeaxanthin, α-carotene and β-carotene from RRIM600 and PB235 latex samples are shown in Figure 3.2.5.1.

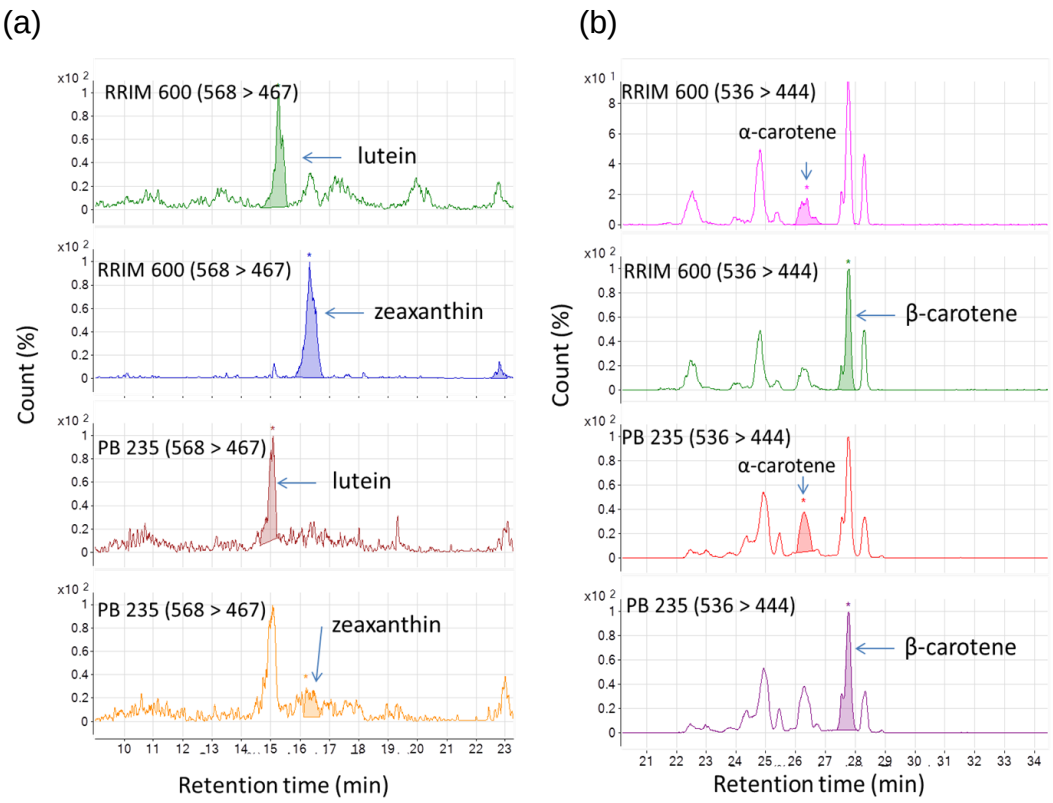


Figure 3.2.5.1. The MRM chromatogram from the RRIM600 and PB235 latex samples. As identical patterns were obtained for all transitions, only MRM chromatogram for $[M-92]^+$ ion products were shown.
(a) Chromatogram generated from precursor ion *m/z*=568 and product ion *m/z*=467 to detect lutein and zeaxanthin. The detection of zeaxanthin in PB235 latex was found to overlap with lutein signal. However, due to differential retention time of both compounds, zeaxanthin could be identified from the sample.
(b) Chromatogram generated from precursor ion *m/z*=536 and product ion *m/z*=444 for identification of carotene isomers. Multiple peaks observed for the MRM chromatograms were due to the multiple carotene isomers present in *Hevea* and in this work, focus was given to α-carotene and β-carotene.

Previously, ten peaks were observed from the HPLC separation of the RRIM600 and PB235 latex extracts. The identical ten peaks were also identified from HPLC-DAD-MRM. For an overview of the relative amount of carotenoid in the latex of RRIM600 and PB235, the corresponding percentage was calculated based on the area of peaks generated from HPLC-DAD-MRM. The percentage of carotenoid composition is summarised in Table 3.2.5.1. β-carotene was the major carotenoid in both latex samples, representing 37.07% and 38.12% of total carotenoids respectively. This was followed by α-carotene, which consisted of 5.42% of total carotenoids in RRIM600 latex and 7.01%% of that of the PB235 latex (total percentages of known and unknown carotenoids = 100%). Zeaxanthin contributed 7.11% and 4.68% of RRIM600 and PB235 latex total carotenoid contents. The lutein percentage was found to be the lowest of the total carotenoids, with 2.20% and 1.39% respectively for RRIM600 and PB235 latex samples.

Another aspect of this study was to quantify lutein, zeaxanthin, α-carotene and β-carotene through DAD-acquired data. The targeted quantification was carried out on non-saponified latex samples. This was based on the observation that non-saponified and saponified HPLC chromatographic profiles are similar to each other. The calibration curves calculated based on the peak area of carotenoid standards are summarised in Table 3.2.5.2. The absolute levels of lutein, zeaxanthin, α-carotene and β-carotene are listed in Table 3.2.5.3. It has been observed that all targeted carotenoids amounts were consistently higher in PB235 latex (4.3 average fold difference) compared to that of RRIM600 latex. β-carotene was found to be the major carotenoid in latex, at 7.8 ug/g in PB235 and 1.9 ug/g in RRIM600.

Table 3.2.5.1: Relative amount (% peak of total peak area) of carotenoid compounds found in RRIM600 (*n*=6) and PB235 (n=4) latex samples.

		Peak 1	Peak 2	Peak 3	Peak 4	Peak 5	Peak 6	Peak 7	Peak 8	Peak 9	Peak 10
		Neoxanthin	Unknown A (HPLC peak 2)	Lutein	Zeaxanthin	β-cryptoxanthin	Unknown B (HPLC peak 6)	Unknown C (HPLC peak 7)	α-carotene	β-carotene	Unknown D (peak 10)
RRIM600	**Range**	9 .00-13.29	2.00 - 4.37	1.30-2.66	3.99 - 10.12	0.88 – 1.67	8.16 – 10.32	8.00 – 8.57	4.97 – 5.98	35.98 – 38.27	13.96 – 14.94
	Mean	11.96	2.98	2.02	7.11	1.43	9.36	8.33	5.42	37.07	14.32
	SD	1.69	0.93	0.66	2.95	0.28	1.05	0.25	0.48	0.92	0.36
PB235	**Range**	1.50 – 2.25	0.72 - 1.9	0.60 – 3.24	2.43 – 5.92	0.41 – 0.77	15.81 – 20.35	5.51 – 7.64	5.51 – 7.64	37.01 – 40.10	13.14 – 13.38
	Mean	2.09	1.13	1.39	4.68	0.56	18.65	13.11	7.01	38.12	13.27
	SD	0.62	0.53	1.24	1.54	0.15	3.78	2.22	1.01	1.41	0.10

Table 3.2.5.2: Calibration curve parameters generated from known concentration (between 0.39 – 100 µ*M*) of carotenoids standards.

Carotenoids	Curve equation	Determination coefficient (r^2)	Limit of detection (µ*M*)	Limit of quantification (µ*M*)
Lutein	72.706*x* - 47.523	0.9989	0.250	0.833
Zeaxanthin	47.35*x* - 35.701	0.9995	0.469	1.564
α-carotene	303.32*x* - 220.72	0.9977	0.472	1.573
β-carotene	108.83*x* - 202.27	0.9889	0.417	1.389

Table 3.2.5.3: Absolute concentration (ug/g) of four major carotenoids in the latex of RRIM600 (*n*=6) and PB235 (n=4) (*p*-value=0.05).

		Lutein	Zeaxanthin	α-carotene	β-carotene
RRIM600	Mean	0.8	0.6	0.7	1.9
	SD	0.0447	0.0007	0.0001	0.0005
PB235	Mean	1.2	0.8	0.9	7.8
	SD	0.0064	0.1596	0.1487	3.4057

3.3. Discussion

Carotenoid is one of the non-rubber isoprenoids found in *Hevea* latex. Even with multiple reports inferring the role of carotenoids in the darkening of the processed latex, there is surprisingly little information about *Hevea* latex carotenoid in the literature. On the other hand, the genetic, the proteomic and the metabolite regulation of carotenoid synthesis have been thoroughly documented in other plant species such as *Arabidopsis*, tomato and potato (Ruiz-Sola and Rodríguez-Concepción, 2012, Liu et al., 2015, Campbell et al., 2014, Enfissi et al., 2017). The knowledge of carotenoid biosynthesis in these plants has been partly driven by the genomic and transcriptomic resources available (for example, *Arabidopsis*) and partly due to the beneficial effects of carotenoids as antioxidants for food consumption (in tomato, potato and rice).

The paucity of carotenoid information from *Hevea* latex is not unexpected. This is due to the technical difficulties in handling the fresh latex samples. *Hevea* tree plantations are mostly operated in tropical or sub-tropical countries and hence fresh latex supply mainly originates from such areas. Once harvested, latex has to be transported from its collection site to the lab and a long transportation procedure will increase the chance of latex becoming irreversibly coagulated. To ensure the latex samples remain fluid, ammonia solution has to be added as a preservation agent. However, ammonia addition may change the compositions of latex metabolites. Therefore, latex samples used in this study were collected and extracted immediately. Due to the complexity of the sample collection and limited access to latex supply, every step involving metabolite extraction, latex metabolite separation, identification and quantification has to be optimised. Such practical issues have to be

considered whilst optimising the extraction method. Despite the limitations in handling the plant materials, carotenoid determination and quantification analysis has been successfully performed on fresh latex samples, based on the HPLC separation by Fraser et al 2000. The method was chosen as it has been applied successfully in separating and identifying 25 classes of carotenoids in a single assay, using extraction products from a wide variety of plant sources such as Arabidopsis, tomato and potato (Fraser et al., 2000, Campbell et al., 2010).

In this study, the optimisations were aimed first to confirm the best working solvent to extract carotenoids from *Hevea* latex and secondly to increase efficiency of the extraction process by comparing extraction recovery from fresh and coagulated *Hevea* latex samples. Generally, non-polar metabolites such as carotenoids can be routinely extracted using organic solvents such as acetone, tetrahydrofuran, petroleum ether or in combination of the solvents (Kao et al., 2012, Inbaraj et al., 2008, Lashbrooke et al., 2010). However, it was observed that the rubber particles in *Hevea* latex would solubilise in some organic solvents. Indeed, the solubilised rubber particles which were observed during total lipid extraction had caused co-extraction of high portion of polyisoprene and this led to the misrepresentation of the non-rubber component (Liengprayoon et al., 2008). The occurrence of solubilised rubber must be circumvented as it will affect the accuracy of spectrometry absorbance and hence introduce bias during carotenoid quantification. Acetone has been demonstrated to efficiently extract carotenoid from latex samples without solubilising the rubber particles. Aside from determining a suitable solvent, the efficiency of carotenoid extraction from fluid and coagulated latex was also evaluated. The extraction recovery was represented by the extraction

recovery of a spiked internal standard (*trans*-β-Apo-8′-carotenal) into the latex samples. The discrepancy in the spiked standard recovery between fluid and coagulated latex samples was due to the carotenoid compound becoming entrapped in the tightly coagulated latex matrix. The lower recovery of metabolite extraction from coagulated latex compared to that of fresh latex was also observed by Liengprayoon et al (2013) during total lipid extraction.

It was observed that the carotenoid contents in both yellow and white latex was dominated by β-carotene. Other carotenoids, namely neoxanthin, lutein, zeaxanthin, β-cryptoxanthin and α-carotene have also been identified from the latex samples. The occurrence of β-carotene in *Hevea* latex has been reported before (Sakdapipanich, 2006). β-carotene is one of the plant pigments that gives a yellow-orange hue to plant tissues. β-carotene is a main contributor of colour to other plant tissues such as roots in the yellow-flesh sweet potato (Tanaka et al., 2017), fruit flesh in apricot varieties (Ruiz et al., 2005) and maize kernel colour (Wong et al., 2004). Additionally, in cauliflower, an elevated β -carotene accumulation through genetic manipulation of genes related to carotenoid formation can change the curd's colour phenotype (from white to yellow-orange) (Paolillo et al., 2004, Lu et al., 2006). In *Hevea* latex, carotenoids were accumulated within FW particles (Gomez and Samsidar, 1989). Indeed, carotenoid accumulation (together with polyphenol oxidases) has been used as a reliable indicator of FW particles (D'Auzac and Jacob, 1989). Therefore, it could be hypothesised that a higher accumulation of β-carotene could be a determinant of the *Hevea* latex colour. The reconciliation of carotenoids accumulation generated through HPLC-DAD-MRM with expression of transcripts involved in the biosynthetic steps of carotenoid will provide further support for the above hypothesis. For instance, experimental evidence has

indicated that the accumulation of β-carotene is positively correlated to the expression of gene encoding phytoene synthase (PSY) in many plant species (Gady et al., 2012, Kim and DellaPenna, 2006, Li et al., 2001). PSY mediated the first carotenoid-committed step, producing ζ-carotene (refer to Figure 3.1.1.1, section 3.1). Therefore, in the subsequent transcriptome analysis of the RRIM600 and PB235 latex samples, it is highly likely that *PSY* may show high expression level.

Aside from contributing to latex colour through the accumulation of β-carotene, carotenoids in *Hevea* latex may play a role in the formation of latex phytohormone biosynthesis pathways. For example, it has been established that the biosynthesis of abscisic acid (ABA), involves derivatives of zeaxanthin as precursors (Liotenberg et al., 1999). Oxidation of the zeaxanthin derivatives with abscisic aldehyde leads to the formation of abscisic acid. It was reported that the formation of zeaxanthin to violaxanthin, mediated by zeaxanthin epoxidase was rate limiting for ABA biosynthesis (Marin et al., 1996). In this study, zeaxanthin and its isomer, lutein were found to be among the major latex carotenoids. It is likely that zeaxanthin in latex is involved in ABA formation. In addition, ABA receptor *PYL* gene members have been identified from the *Hevea* draft genome and their corresponding transcripts were found to be highly expressed in laticifer tissue (Guo et al., 2017). These genes were reported to be involved in the ABA-signalling pathway in modulating plant development and stress response (Finkelstein et al., 2002, Park et al., 2009).

DRC has been established as one of the metrics used for the screening of rubber trees that possess high levels of latex production (Eng et al., 2001, Yip, 1990). Based on a study of 28 rubber tree clones over a period of one year, it has been proposed that rubber tree clones generating DRC value higher than

41% can be categorised as a high-yielding (Yip, 1990, Eng et al., 2001, Ong, 2000). In this study, it was observed that the RRIM600 latex contained a higher rubber content compared to that of PB235. However, it was noted that the discrepancy of the DRC between these two genotypes were quite small. Indeed, in another study, Chow et al (2012) observed comparable DRC values in both PB235 and RRIM600 latex samples. Nevertheless, both PB235 and RRIM600 can be categorised as high-yielding *Hevea* tree genotypes as the DRC values generated in this study were observed to be more than 41%. Additionally, a high DRC value is also linked to higher expression of the isoprenoid biosynthetic pathway transcripts. Increased expression level of genes such as hydroxy-methylglutaryl coenzyme A reductase (*HMGR*), rubber elongation factor (*REF*), small rubber particle protein (*SRPP*) and *cis*-prenyltransferase were documented in the *Hevea* latex with high DRC clones (Suwanmanee et al., 2013, Ji et al., 1993, Collins-Silva et al., 2012, Yamashita et al., 2016).

In conclusion, the HPLC-DAD-MRM method has detected six carotenoids from the latex of *Hevea brasiliensis.* The method is highly sensitive and allows small volume of the latex samples to be used for carotenoid identification and quantification. From the HPLC-DAD-MRM analysis, four major carotenoids were detected in the latex of *Hevea brasiliensis* and the accumulation of carotenoid in PB235 is higher than that of RRIM600 latex.

Chapter 4

Development and optimisation of an analytical method for the profiling of isoprenoid intermediates in the latex of *Hevea brasiliensis*

4.1.1. Brief introduction

Isoprenoids are the largest group of plant natural compounds, derived from isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) intermediates. The main isoprenoid product in the latex of *Hevea brasiliensis* is rubber. Apart from rubber, other isoprenoid products such as carotenoids, dolichol, tocopherol, tocotrienols have been reported in rubber (Hasma and Subramaniam, 1986, D'Auzac and Jacob, 1989, Dunphy et al., 1965, Whittle et al., 1967, Phatthiya et al., 2007, Tateyama et al., 1999). Although these non-rubber isoprenoids occur in in relatively smaller amounts in latex, they exert significant role in in rubber physical property.

As mentioned previously in Chapter 1, section 1.3, MVA is assumed to be the main source of IPP whereas the MEP pathway supplies IPP for the plastidic isoprenoids such carotenoids (Archer et al., 1963, Lichtenthaler, 1999). Plastidic IPP is involved in the formation of GGPP (C20) whereas the cytosolic IPP is utilised in the formation of FPP (C15) (Vranová et al., 2013). Both FPP and GGPP provide the basis for the formation of rubber and carotenoid respectively (Cornish and Siler, 1995, Hirschberg, 2001). There is a hypothesis that suggests that the IPP generated from the MEP pathway may be transported from the plastid into cytosol and may be involved in rubber formation (Chow et al., 2012). However, no experimental evidence currently exists to ascertain the cross-talk of the IPP in the latex of *Hevea brasiliensis* for rubber formation. In contrast, such cross-talk has been reported in other plant species such as in monoterpene biosynthesis in spiked lavender (Mendoza-Poudereux et al., 2015), in the formation of sesquiterpenes and monoterpenes of tea leaves (Xu et al., 2018) and sesquiterpenes in stevia leaves (Wölwer-

Rieck et al., 2014). Despite some accumulating evidence in these plant species, the detailed mechanisms that underlie the cross-talk of both biosynthetic routes are not fully understood. To gain information on the role of the MEP pathway in rubber formation, it would be beneficial to know the levels of the MEP intermediates and their relationship to rubber accumulation.

4.1.2. Metabolite profiling methods

An efficient and robust analytical method needs to be employed so that an accurate complete overview of the metabolite state could be drawn. Multiple approaches have been successfully used to dissect metabolite profiles in plants. This includes nuclear magnetic resonance (NMR) and mass spectrometry (MS) -based approaches such as gas chromatography mass spectrometry (GC-MS) or liquid chromatography mass spectrometry (LC- MS).

NMR offers unequivocally accurate identification of metabolites (Markley et al., 2017). However, the technique requires a large amount of metabolite extracts and is less sensitive in quantifying the analytes (Giraudeau, 2017). On the other hand, GC-MS and LC-MS approaches utilise smaller amounts of analytes and can detect analytes with a higher degree of sensitivity. GC-MS operates by separating analytes based on their relative vapour pressure. This requires derivatisation of the non-volatile analytes so that they are more amenable for GC-MS fractionation. However, such sample pre-treatment may degrade or modify their chemical and/or physical structures. In contrast, LC-MS techniques can be successfully applied to limited amount of starting material and does not require sample volatilisation. Due to the limitations of NMR which demands abundant starting materials and GC-MS which needs the transformation of the starting materials into volatile derivatives, an LC-MS

technique was used in this study to detect targeted isoprenoid compounds from the latex of *Hevea brasiliensis.* As mentioned previously in Chapter 1 section 1.7, latex collection and extraction has to be performed *in-situ* to avoid latex coagulation. Therefore, an analytical method that uses a small amount of starting material is more suitable for the metabolite profiling of *Hevea brasiliensis* latex.

At the moment, no generic protocol been developed to extract isoprenoid intermediates from the latex of *Hevea brasiliensis*, followed by efficient separation and accurate identification of the extraction products using LC-MS. Recently, Zhang et al. (2018) has reported an analytical method to quantify FPP and DMAPP in the latex of *Hevea brasiliensis.* However, their method does not provide a protocol to investigate the MEP pathway intermediates. Many reports using LC-MS protocols have been published using tissue extracts from other plant species in analysing isoprenoid intermediates, such as in poplar leaf (Li and Sharkey, 2013), tomato (Balcke et al., 2017), tobacco and rosemary trichomes (Balcke et al., 2014). The selection of LC-MS protocols is dictated by the chemical and physical behaviours of the compound of interest. The molecular weight of isoprenoid intermediates targeted in this study ranges from 214 – 450 Da, this range is universally classified as low molecular weight compounds or small metabolites. These compounds are involved in varied biological reactions as either substrates or products of biochemical reactions in living organisms. Not only is the molecular weight of targeted compounds in this study classified as low, the water-soluble isoprenoid compounds are also of varied chemical structures and physical behaviours. For example, the polarity of these compounds is varied, depending on their hydrocarbon backbone length. This is because whilst carrying negatively-charged diphosphate groups,

isoprenoid intermediates also consist of repeats of covalently-linked non-polar hydrocarbon chains. Although the hydrocarbon chains are hydrophobic, this property is often masked by the charged functional groups within their molecule structure. Therefore small isoprenoid intermediates such as IPP will exhibit hydrophilic, polar properties due to their shorter carbon chain whilst longer isoprenoid intermediates such as FPP (C_{15}) and GGPP (C_{20}) tend to be less polar compared to the smaller isoprenoid intermediates (Balcke et al., 2014).

A method known as hydrophilic interaction liquid chromatography (HILIC) (Alpert, 1990) is currently the most used LC method in efforts to separate highly polar, small metabolites (Tolstikov and Fiehn, 2002, Bajad et al., 2006, Hsieh, 2008, Appulage and Schug, 2017, Naz et al., 2017). Currently, the HILIC approach has been reported to be successfully applied in characterising metabolites with a wide range of polarity from plants (Gika et al., 2012, Liu and Rochfort, 2013, Li and Sharkey, 2013, Ito et al., 2014, Salem et al., 2016).

The HILIC technique not only provides a seamless movement of analytes from a chromatographic column to the fragmentation phase for subsequent mass spectrometry, it also provides a better separation of the analytes during the chromatographic analysis. This is because the HILIC method uses an ionic additive in an aqueous solvent to control the pH and ionic strength of the phase. This will assist the partitioning of the compound of interest into either polar or water-enriched layers that are formed within the mobile phase during the chromatographic run. Simultaneously, the compounds will also interact with the stationary phase of the column through adsorption. The partition and adsorption of compounds of interest between mobile and stationary phases will create selectivity in the HILIC separation and hence, governs retention during the chromatographic run.

However, due to these complex separation mechanisms, if the chromatographic conditions are not properly optimised, HILIC can give low reproducibility and poor resolution of HPLC peaks. For example, broadening of the peaks, peak tailing or peak splitting are often observed in the HILIC separation (Wernisch and Pennathur, 2016). This will hinder the subsequent downstream analysis, such as peak integration and identification. For these reasons, it is imperative to optimise chromatographic conditions used for the isoprenoid intermediate profiling. Therefore, experiments were carried-out with the aim of determining the best chromatographic condition and the distribution of chromatogram peaks during the HILIC run.

Apart from lending sensitivity in separating the targeted polar isoprenoid intermediates, HILIC has been chosen for this study as its aqueous mobile phase allows consecutive ion fragmentation of the analytes for compound identification through mass spectrometry (MS) techniques. MS allows determination of molecular masses of biomolecules such as phospholipids, sugars, proteins and polymers (Domingues et al., 2008, Cha and Yeung, 2007, Lung and Liu, 2015). Not only will it assist in the determination of the mass of different compounds, MS can also facilitate the study of molecular species and hence, aiding in the qualitative identification of compounds of interest. After the chromatographic separation, an analyte can be channelled to the MS apparatus. Subsequently, the analyte will be fragmented and sorted on the basis of their ion mass (or mass-to-ion ratio, *m/z*) when the fragmentised ions are accelerated/deflected through series of high-voltage electrodes.

4.1.3. Aims

Information on the levels of isoprenoid metabolites in latex is highly relevant to gaining a deeper understanding of the regulation of isoprenoid biosynthesis. Indeed, the levels of the isoprenoid end product depends on the supply of its precursors, as has been demonstrated in the synthesis of *Hevea*'s rubber *in vitro* (Archer et al., 1961), the formations of Arabidopsis' sterols and peppermint's dolichols (Lange, 2015, Akhtar et al., 2017) and in the synthesis of carotenoids in *Adonis aestivalis* (Botella-Pavía et al., 2004). Since the changes in the intermediate contents will impact the levels of isoprenoid products, it is highly advantageous to elucidate the regulatory mechanism of the isoprenoid biosynthesis pathway. By measuring these intermediates, information about the isoprenoid metabolic pathway activities can be inferred. This includes the co-response between the intermediate precursors and isoprenoid end products. Therefore, by integrating transcriptomic and metabolomic data, a deeper understanding of isoprenoid biosynthesis and the availability of intermediates in the metabolic pool for the production of isoprenoid end products can be gained. Information about the relative IPP utilisation in latex may help in developing strategies to enhance IPP supply directed towards rubber biosynthesis whilst reducing the competition from non-rubber isoprenoids.

In an effort to survey the profile of isoprenoid intermediates in *Hevea* latex, a HILIC-MS/MS method was developed by firstly, modifying HILIC chromatogram conditions and secondly, through the determination of characteristic MS ion fragment patterns of the analytes. This was attained when a second level of MS was performed through electrospray ionisation, to enable assignment of an accurate structure to the product ions. The optimised method was then applied to separate and identity isoprenoid intermediates from potato

extracts, which served as an abundant source of plant material for the method development. Subsequently, isoprenoid intermediates extracted from the latex of *Hevea brasiliensis* genotypes (RRIM600 and PB235) were separated and identified using the developed HILIC-MS/MS method.

4.2. Results

4.2.1. HILIC optimisation

To perform the optimisation, the HILIC-MS/MS technique reported by Li and Sharkey (2013) was used as a starting point for setting up the running conditions. During the optimisation of the HILIC method, a mixture of analytical standards containing IPP, DMAPP, GPP, GGPP, FPP, DXP and MEP (100 µM each) was separated as described in Chapter 2, Materials and Methods section 2.12. The mixture of analytical standards used in this study provides a representative of a complex sample extract.

Firstly, the optimisation was performed by evaluating two columns with different stationary phase packing materials (Table 4.2.1.1). The best column was selected based on the ability to produce an identifiable chromatographic peak, with stable retention time. In summary, the Accucore 150 Amide HILIC packing material consisting of partially porous silica particles. This contributes to a lower column back pressure. Having a relatively low column back pressure will ensure a long life span of the HILIC column. Additionally, the amide functional group within the silica particles will bind hydroxyl groups present in analytes and this provides a robust system for the separation of hydrophilic biomolecules. On the other hand, SeQuant® ZIC®-*p*HILIC is packed with polymeric particles, with attached sulfobetaine-groups. This promotes the

formation of zwitterionic conditions for the stationary phase which then could accommodate a wide spectrum of pH changes, hence facilitate the separation selectivity. Whilst assessing the stationary phase aspect, the mobile phase composition reported by Li and Sharkey (2013) was applied. The mobile phase consisted of ammonium acetate buffer 10 mmol/L, pH (Solvent A) and acetonitrile (100%, vol/vol) (Solvent B). The gradient was initiated with 20% Solvent A and 80% Solvent B for the first 2 minutes. Subsequently, a linear gradient elution was applied (30% -60% Solvent A) in 13 minutes before the re-equilibration of the column in 20% Solvent A and 80% Solvent B was applied for 45 minutes.

Table 4.2.1.1: HILIC columns and its packing materials info

Column	**Particles in the packing material**	**Functional group of the stationary phase**	**Dimension**	**Particle size (µm)**
Accucore 150 Amide HILIC	silica	amide	100x2.1 mm	2.6
SeQuant® ZIC®-*p*HILIC	polymer	sulfobetaine-group	150x2.1 mm	5

The HILIC chromatograms of seven isoprenoid intermediates separated through two different columns are shown in Figure 4.2.1.1. When operating on the Accucore 150 Amide HILIC column, it was observed that the reproducibility of the HPLC spectra was low. This is because in the first attempt to separate the standard mixtures, split peaks were observed for IPP/DMAPP, MEP and DXP (Figure 4.2.1.1a). Since the analytes consisted of only analytical standards, only a single peak for each analyte is expected. The split peak problem was resolved when the run was repeated (Figure 4.2.1.1b), but with a decreased concentration of the analytes (from 300 µM to 100 µM). This indicated that split peaks from the HILIC run were attributed to overloading of

the samples and hence the acquired HPLC signals were saturated. Although the split peaks issue was resolved, another issue encountered when using Accucore 150 Amide HILIC was that the HPLC spectra showed tailing at the end of the peaks. The tailing peak phenomena is often associated with poor peak shape during chromatographic separation. The retention distribution of the analytes showed that the compounds with lower molecular weight (such as IPP, DMAPP, MEP and DXP) were retained longer in the column compared to the compounds with higher molecular weight (GGPP, FPP and GPP). IPP and DMAPP were found to be co-eluted from the HILIC column. This is because the compounds are isomers and share the same molecular weight.

Another observation made for the HILIC separation using the Accucore 150 Amide column was the retention shift between different runs (± 4 minutes). On the other hand, when the analytes were separated using SeQuant® ZIC®-*p*HILIC column, the peak shapes for the analytes were generally improved, as the peak shapes were observed to be narrower (Figure 4.2.1.2). However, similar issues observed during the HPLC run using the Accucore 150 Amide column were also recorded using the SeQuant® ZIC®-*p*HILIC column. This can be seen from the peak tailing generated from GPP and FPP. Additionally, double peaks were observed for GGPP. Clearly, the broadening of the peaks for some of the compounds was not improved by just using the SeQuant® ZIC®-*p*HILIC column with mobile phase parameters set according to the HILIC technique reported by Li and Sharkey (2013).

Although HILIC separation for both columns exhibited poor peak shapes for some of the compounds, the peak shapes could be further improved through the manipulation of mobile phase polarity. This involved evaluation of parameters of the mobile phase such as varied pH ranges. However, the Accucore 150 Amide

column could not function well in a pH of the mobile phase higher than 9 and it operated with lower concentration of mobile phase solution (less than 5mM) (Thermo-Fisher Scientific, 2018). On the other hand, the SeQuant® ZIC®-*p*HILIC column is more robust, as it can accommodate a broader range of pH for the mobile phase. Therefore, based on the pH range characteristics, the SeQuant® ZIC®-*p*HILIC column was preferred over the silica-based Accucore 150 Amide column for the subsequent mobile phase parameter optimisation.

Similar to the evaluation of the HILIC columns, the main purpose in determining the parameters of the mobile phase for HILIC separation was to produce good peaks shapes for the subsequent downstream metabolite identification. For the assessment of mobile phase conditions, the SeQuant® ZIC®-*p*HILIC column was utilised for the separation of seven isoprenoid intermediate standards. A total of three mobile phase parameters were evaluated; namely gradient of the mobile phase composition, the flow rate of the mobile phase movement within the HILIC column and pH of the aqueous solution that formed part of the mobile phase (Table 4.2.1.2).

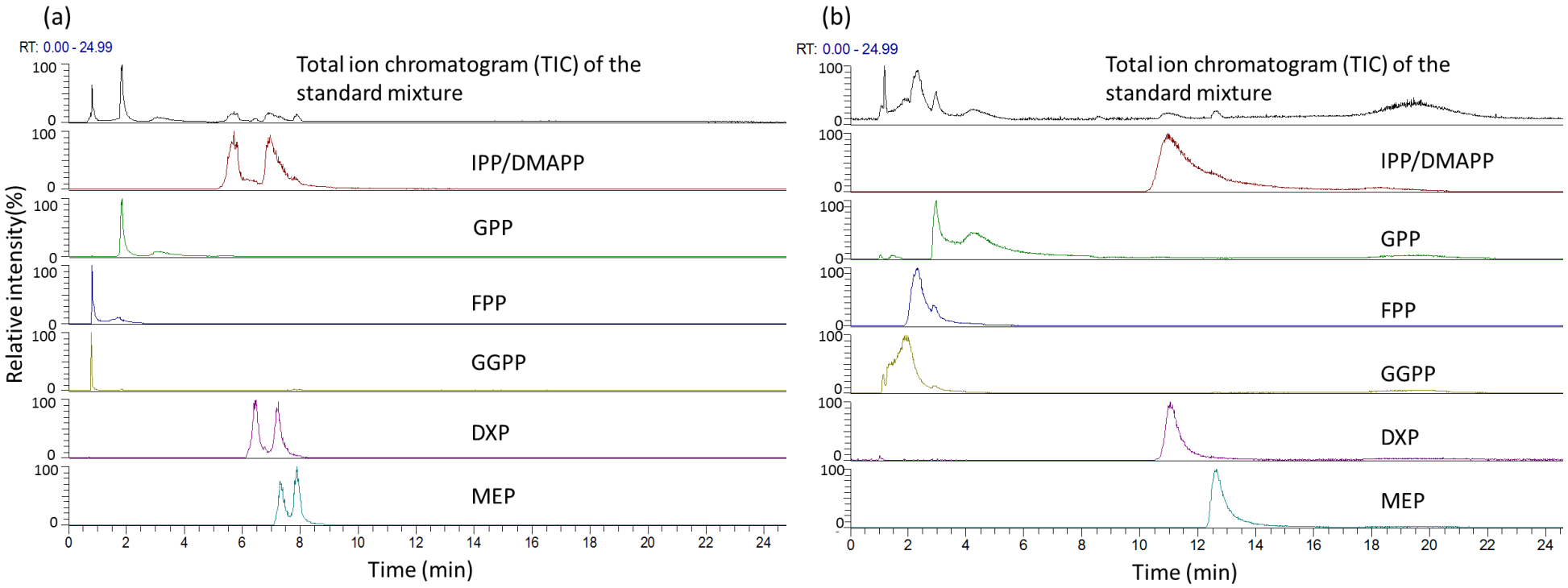


Figure 4.2.1.1. (a) The fractionation of seven isoprenoid intermediate standards from Accucore 150 Amide column. The HILIC separation for IPP/DMAPP, DXP and MEP gave rise to split peaks. However, GPP, FPP and GGPP showed good HILIC peaks. (b) The same fractionation of seven isoprenoid intermediate standards were repeated, with a lower concentration of the injected analytes (decreased from 300 μM to 100 μM). The splitting peaks issue was resolved but the analyte separation gave rise to broad and tailing peaks.

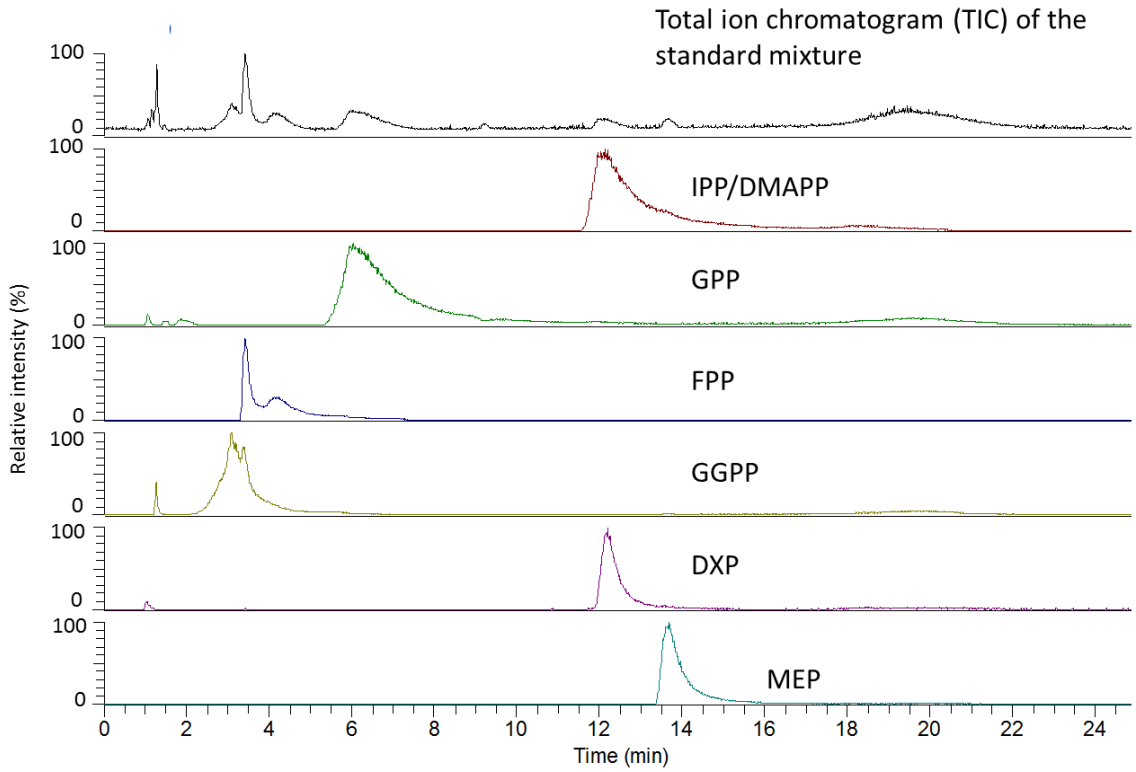


Figure 4.2.1.2. HILIC separation of isoprenoid intermediate standards using SeQuant® ZIC®-*p*HILIC column

Table 4.2.1.2. Mobile phase parameters evaluated during HILIC method development.

Mobile phase condition	Flow rate tested	Ammonium acetate pH
Gradient 1	200 µl/min and 400 µl/min	pH 9
Gradient 2	400 µl/min and 500 µl/min	
Gradient 3	200 µl/min and 400 µl/min	
Gradient 4	200 µl/min and 400 µl/min	
Gradient 1	200 µl/min and 400 µl/min	pH 10
Gradient 2	400 µl/min and 500 µl/min	
Gradient 3	200 µl/min and 400 µl/min	
Gradient 4	200 µl/min and 400 µl/min	
Gradient 1	200 µl/min and 400 µl/min	pH 11
Gradient 2	400 µl/min and 500 µl/min	
Gradient 3	200 µl/min and 400 µl/min	
Gradient 4	200 µl/min and 400 µl/min	

The mobile phase for HILIC separation is usually composed of aqueous solution (usually water or buffer such as ammonium acetate) and acetonitrile (100% vol/vol). The elution of compounds of interest during HILIC separation is attributed to the rising polarity of the mobile phase. Generally, under HILIC separation, when a higher percentage of aqueous solution was utilised, a higher retention of small and polar compounds was obtained. Therefore, by modifying the gradient of mobile phase composition, the separation of the targeted compound can be optimised.

Three different mobile phase gradients were assessed (Table 4.2.1.3). Gradient 1 represented the mobile phase composition that was previously reported by Li and Sharkey (2013). Gradient 2 was modified from Gradient 1, whereby a shallower increase of aqueous solution (from 30% to 40%) was applied for 9 minutes before re-equilibration of the HILIC column. Gradient 3 was formed by increasing the time for the gradient of aqueous solution (from 30% to 40%), for 17 minutes before the HILIC column was returned to the initial condition of 20% aqueous solution. Additionally, Gradient 4 has the longest running time, due to the increased time for the gradient setting (an increment of Solution A percentage from 30% to 40%) of HILIC mobile phase. The modified gradients were tuned so that it would provide a different selectivity for polar metabolites.

Table 4.2.1.3: Gradients for HILIC mobile phase tested during method development. A: Ammonium acetate buffer 10 mM pH 10, B: Acetonitrile (100%; vol/vol).

Gradient 1		
Time (min)	**A (%)**	**B (%)**
0	20	80
2	20	80
2.01	30	70
6.5	30	70
7.5	60	40
10	60	40
10.01	20	80
15	20	80
Gradient 2		
Time (min)	**A (%)**	**B (%)**
0	20	80
2	20	80
3	30	70
7.5	30	70
8.5	60	40
11	60	40
12	20	80
17	20	80
Gradient 3		
Time (min)	**A (%)**	**B (%)**
0	20	80
4	20	80
6	30	70
15	30	70
17	60	40
22	60	40
24	20	80
34	20	80
Gradient 4		
Time (min)	**A (%)**	**B (%)**
0	20	80
8	20	80
12	30	70
30	30	70
34	60	40
44	60	40
48	20	80
60	20	80

Following the mobile phase gradient evaluation, Gradient 1 gave good separation for FPP, GGPP, DXP and MEP (Figure 4.2.1.3). However, it was noted that tailing and shouldering of peaks for these compounds were also generated using this gradient. Additionally, DMAPP and IPP compounds were co-eluted from the HILIC column. All detected intermediates were eluted between 1.01 and 12.59 minutes. On the other hand, Gradient 2 resulted in narrower chromatogram peaks for FPP, DXP and MEP (Figure 4.2.1.3). However, FPP and GGPP standards gave broad chromatogram peaks with long tails. Whereas GPP and GGPP still gave rise to tailing peaks, the peak shape was generally improved compared to the Gradient 1 results. IPP and DMAPP compounds were still found to be co-eluted. The overall retention times were found to be shifted in this gradient. This is possibly due to a slight change of polarity in solvent composition and hence, the compounds migrated at a slower rate in the column. In both Gradient 3 and Gradient 4, the peak shapes were distorted, giving rise to poor resolution of the standards. Therefore, in the subsequent method development, Gradient 3 and Gradient 4 were not included.

Next, the mobile phase flow rate was evaluated using 0.2 ml/min, 0.4 ml/min and 0.5 ml/min. Injections of standard mixtures into the HILIC column were performed at each flow rate. The flow rate of the mobile phase volume from the end-to-end of the HILIC column contributes to the relative peak spacing and shapes. The mobile phase velocity impacts on the chromatogram peak shape and the throughput of the sample. Therefore, retention time for analytes would vary with the flow rate change. The flow rate optimisation showed that the overall retention time shifted when the flow rate increased (Figure 4.2.1.4). However, it was observed that the order of elution for all compounds was consistent in all the tested flow rates (GGPP > FPP > GPP > DXP > MEP > IPP/DMAPP). The peak

shapes were similar in for the tested flow rates, where tailing of GPP, FPP and GGPP. However, it was observed that the peak shapes showed the most improvement when the mobile phase flow rate was set at 0.4 ml/min.

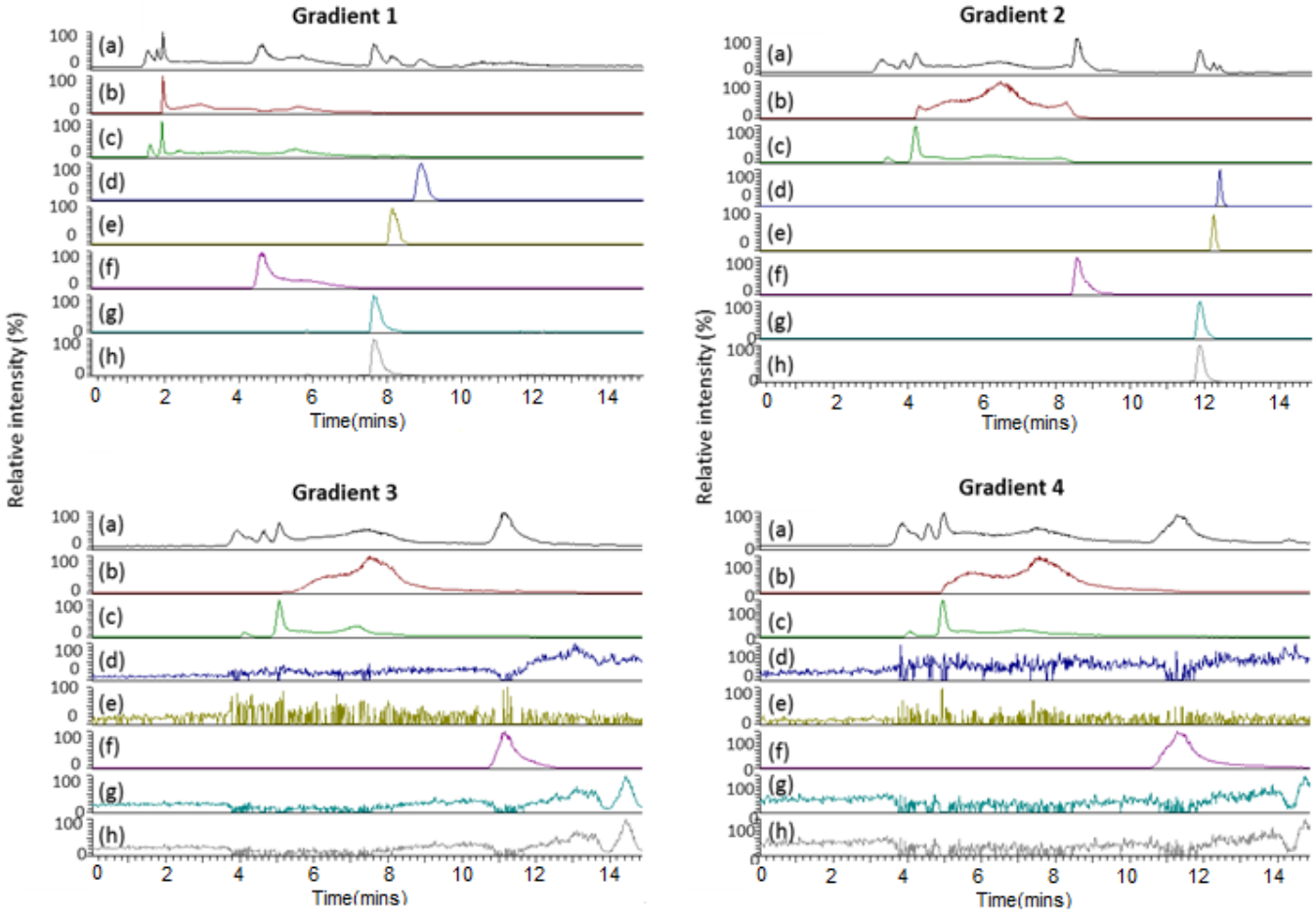


Figure 4.2.1.3: Evaluation of gradient for mobile phase of HILIC for the separation of seven isoprenoid intermediate standards. The chromatograms showed (a) total ion content of the standard mixture; extracted chromatogram for (b) FPP; (c) GGPP; (d) MEP; € DXP; (f) GPP; (g) IPP; and (h) DMAPP. Gradient 2 gave the best separation profile compared to other gradient profiles.

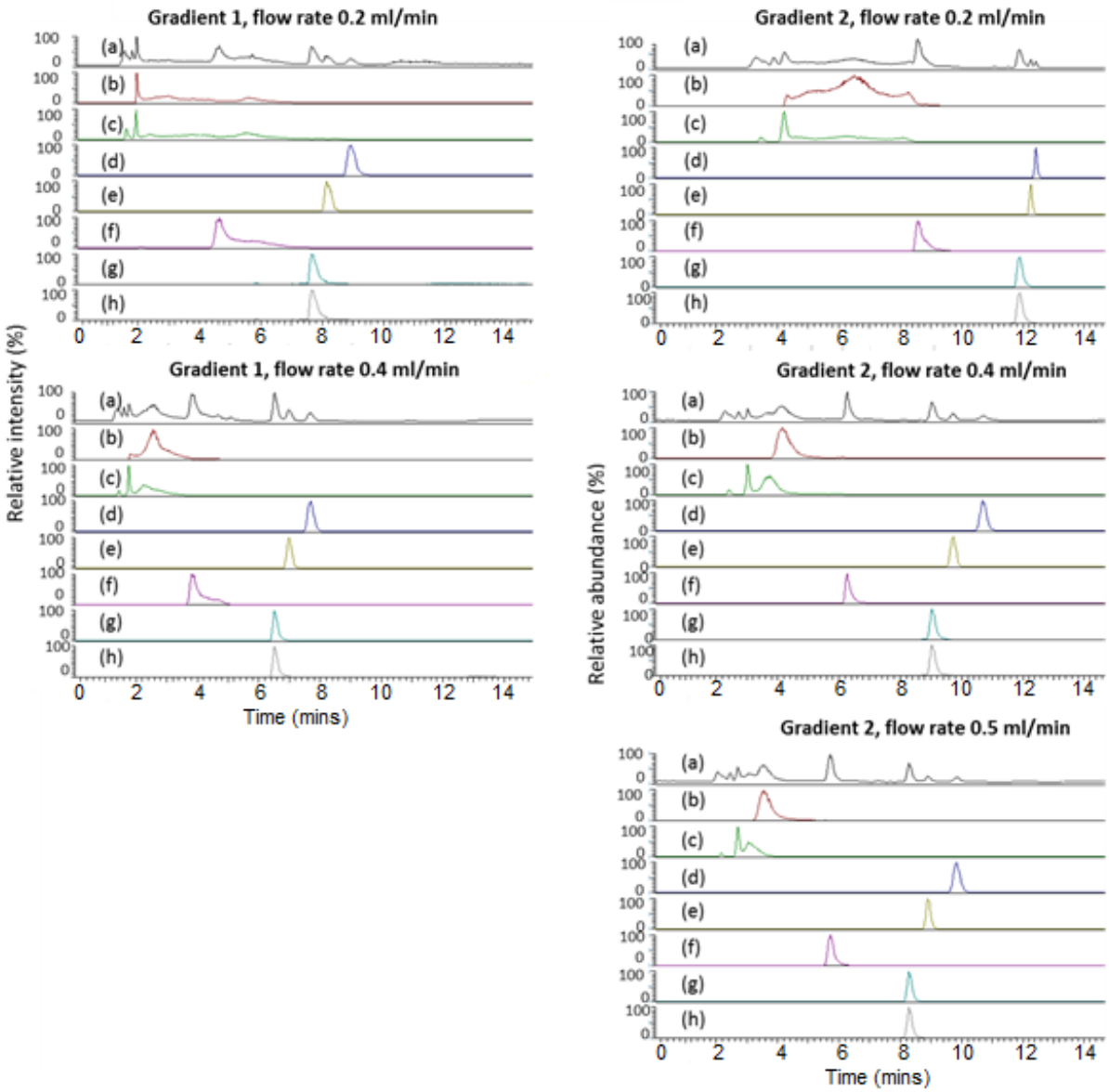


Figure 4.2.1.4: Separation of seven isoprenoid intermediates using flow rates of 0.2ml/min, 0.4 ml/min and 0.5ml/min. The evaluation of flow rates was performed using Gradient 1 separation and Gradient 2 separation. The resulted chromatogram shows (a) total ion content of the standard mixture; extracted chromatogram for (b) FPP; (c) GGPP; (d) MEP; € DXP; (f) GPP; (g) IPP; and (h) DMAPP.

Finally, the effect of the pH of the ammonium acetate buffer, which was used as a part of the mobile phase for the HILIC separation was evaluated. The targeted compounds varied in polarity and size. Therefore, running conditions have to be optimised to accommodate the fractionation of all targeted compounds in a single run. During HILIC separation, small and polar compounds are partitioned into the aqueous layer within the stationary phase whilst the less polar analytes have a shorter retention time as they tend to partition within the solvent layer of the mobile phase. Elevated pH is expected to enhance the polarity of the HILIC mobile phase. Hence, higher pH will increase the retention of polar analytes in the HILIC column. To investigate the effect of different pH values of the ammonium acetate buffer on the chromatogram peaks, a mixture of isoprenoid standards was separated in the HILIC column at a flow rate of 0.4 ml/min using Gradient 2 (which was optimised earlier). The results of the pH evaluation are shown in Figure 4.2.1.5. HILIC separation using ammonium acetate buffer of pH 11 generated the best peak shapes for GGPP, MEP, DXP and GPP. The separation of FPP still gave rise to double chromatogram peaks and IPP still co-eluted with DMAPP.

From the optimisation of the HILIC method, it was concluded that the best working conditions to separate the seven isoprenoid intermediates were by utilising the SeQuant® ZIC®-*p*HILIC polymeric column, with gradient mobile phase (Gradient 2, as optimised previously), aqueous solution adjusted to pH 11 and mobile phase flow rate at 0.4 ml/min. It was observed that two technical issues encountered during HILIC optimisation were 1) difficulty in generating good peak shape for compounds with higher molecular weight (FPP and GGPP) and 2) the co-elution of IPP and DMAPP. Although these issues persisted after the method was optimised, tailing or double peaks of FPP and GGPP could still

be resolved through the derivation of the extracted ion chromatogram (EIC) peak. The EIC peak could be generated from a full mass spectrometry scan of the compounds. Similarly, the identification of FPP and GGPP could be further supported through the ionisation spectra generated by tandem mass spectrometry. On the other hand, due to the co-elution of IPP and DMAPP, the optimised HILIC protocol could not be used to resolve the presence of IPP and DMAPP in a sample. It was not possible to discriminate between these compounds as both IPP and DMAPP are isomers, share similar molecular weight and only differ at the position of the double bond within their carbon backbone.

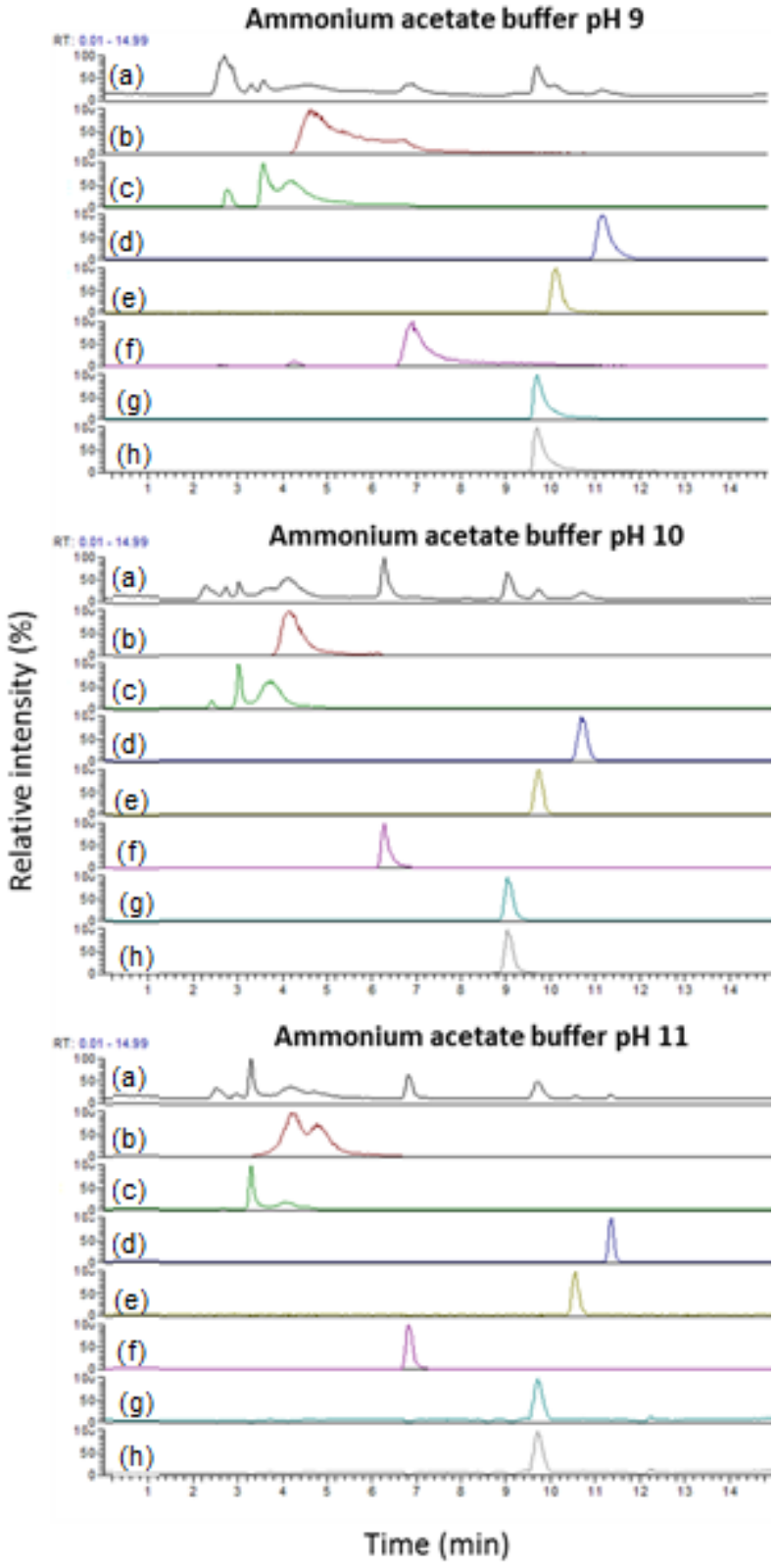


Figure 4.2.1.5. Separation of seven isoprenoid intermediates using ammonium acetate adjusted to pH 9, pH 10 and pH 11. The evaluation of flow rates was performed using Gradient 2 separation and flow rate of 0.4 ml/min. The resulted chromatogram shows (a) total ion content of the standard mixture; extracted chromatogram for (b) FPP; (c) GGPP; (d) MEP; € DXP; (f) GPP; (g) IPP; and (h) DMAPP.

4.2.2. Mass spectrometry

Following the optimisation of the HILIC protocol, it was observed that separating the peaks of isomeric compounds of IPP and DMAPP remained challenging as these compounds share similar chemical structure and the same molecular weight. Likewise, the tailing peaks for GGPP and FPP remained unresolved resulting in difficulties with metabolite identification. To address these technical issues, mass spectrometry (MS) analysis was utilised in conjunction to the developed HILIC protocol. Through MS, information related to the molecular mass of compound of interest could be generated following ionisation of the analytes. The molecular masses corresponding to chromatogram peaks would be used to tentatively assign a chemical structure to the fractionated peak. This framework helps to identify peaks that are not fully resolved following HILIC separation. Indeed, the combination of multiple data from HILIC separation, accurate mass and fragmentation is widely used and accepted to characterise metabolites. Recent applications include combining HILIC technique to MS, to separate and identify polar metabolites from various biological matrices such as plant tissues (Tolstikov and Fiehn, 2002, Ito et al., 2014), animal tissues (Miękus et al., 2017) and microbial extracts (Liu et al., 2016).

In generating mass information from MS, the targeted compound will be subjected to ionisation processes, either by protonation or deprotonation of the analyte. Following the ionisation process, the analyte is generally broken into parent and transition ions. These ions are then propelled through electric fields. The acceleration of the ionised fragments is detected by the mass analyser, which is later derived into a mass-to-charge ratio (*m/z*) value. A full scan of MS

spectra from such reactions allows broad detection of both parent and precursor ions of any compound. By detecting the ion mass of the generated parent ion, accurate mass information for the targeted compounds can be inferred. Additionally, further information regarding chemical structure related to the compounds of interest can be generated when the fragment ions are split during MS/MS. Under MS/MS reactions, fragment ions of interest are selected based on *m/z* values generated during the full scan MS and is passed through a second fragmentation process and the corresponding ion mass detected from this stage can be used to ascribe a possible chemical structure that the *m/z* value originated from.

The main purpose of this experiment was to integrate MS to the developed HILIC method and to identify fragment ions generated through MS for the targeted isoprenoid intermediates. The annotated fragment ions could be used as further supporting evidence in identifying isoprenoid intermediate compounds. The MS experiment was carried-out as described in Chapter 2, Materials and Methods. In this study, the HILIC-fractionated isoprenoid standard mixture was passed through the electrospray ionisation chamber of the Thermo LTQ-Orbitrap XL mass spectrometry system for fragmentation in negative mode. A previous report by Li and Sharkey (2013) concerning some of the isoprenoids (MEP, DXP and DMAPP) indicated that negative mode ionisation gave better fragment ion intensity and therefore, no attempt of ionisation in positive mode was utilised in this study. A data-dependent MS/MS was performed within the Collision Induced Dissociation (CID) chamber in negative mode for the three most abundant ions, defined from the full MS scan event.

The chromatogram peaks from the full scan MS of an isoprenoid mixture is shown in Figure 4.2.2.1. From the peaks, retention times for each targeted

compound were inferred. Additionally, an extracted ion chromatogram (EIC) was generated by monitoring the expected *m/z* for deprotonated compounds ($[M-H]^-$). From Figure 4.2.2.1, it was observed that each MEP and DXP separation and fragmentation gave rise to a single narrow peak. On the other hand, GPP, FPP and GGPP HILIC-MS each exhibited tailing peaks. The monitoring of expected *m/z* for each compound showed the tailing region of the peaks has a *m/z* value corresponding to that of the GPP molecular ion (*m/z* = 381). This indicated that the tailing of the peak could be attributed to either the presence of impurities in the corresponding isoprenoid standards or the degradation of diphosphate moieties that may happen for these isopentenyl diphosphate compounds (Dr William Allwood, personal communication). The HILIC-MS analysis demonstrated that IPP and DMAPP eluted as a single peak. The retention times and *m/z* values are summarised in Table 4.2.2.1.

By monitoring fragment ions between *m/z* 70-1000 during the CID-ionisation, it was observed that the ion species for the fragmented isoprenoid standards were predominantly deprotonated molecular ions $(M-H)^-$ (Figure 4.2.2.2). Dimeric molecular ions $(M+M-H)^-$ and trimeric molecular ions $(M+M+M-H)^-$ were also detected for all compounds, but at a lower intensity level. The MS/MS spectra indicated that all compounds produced three common fragmentation patterns (Table 4.2.2.1). Firstly, each CID-fragmented isoprenoid compounds exhibited a loss of water moiety, producing the $[M-H-H2O]^-$ ion. Secondly, other characteristic fragment ions generated from the isoprenoid fragmentation were at *m/z* 79 and 159. The *m/z* 79 and *m/z* 159 ions are the common fragments produced during the dissociation of isoprenoid diphosphates, due to the loss of the diphosphate ester moiety (Nürenberg and Volmer, 2012, Lange, 2015). In addition, IPP/DMAPP CID-fragmentation

produced the *m/z* 177 ion and DXP fragmentation generated the *m/z* 139 ion. The *m/z* 177 ion was attributed to a phosphate moiety ion, which was cleaved from the compound during MS/MS fragmentation. The *m/z* 139 ion produced from the DXP fragmentation was ascribed to an ether group, generated after the molecular ion was dissociated from the ethyl phosphate moiety.

Unique *m/z* values for ions generated from ESI-fragmentation (in the form of deprotonated molecular ions) and the MS/MS reaction (in the form of an ion that was generated due to the loss of a single water residue) were recorded for GPP, FPP, GGPP, MEP and DXP. However, IPP and DMAPP, eluted as a single peak during HILIC fragmentation and these compounds also shared similar fragmentation patterns. Therefore, IPP and DMAPP could not be measured individually using the HILIC-MS/MS protocol developed in this study. However, the issue of broad peaks for GGPP and FPP (which could not be separated during HILIC optimisation) was resolved through the comparison of the associated peak to their unique MS/MS ion information. Likewise, full scan MS and MS/MS information would also be used to facilitate subsequent identification of MEP and DXP from biological matrices.

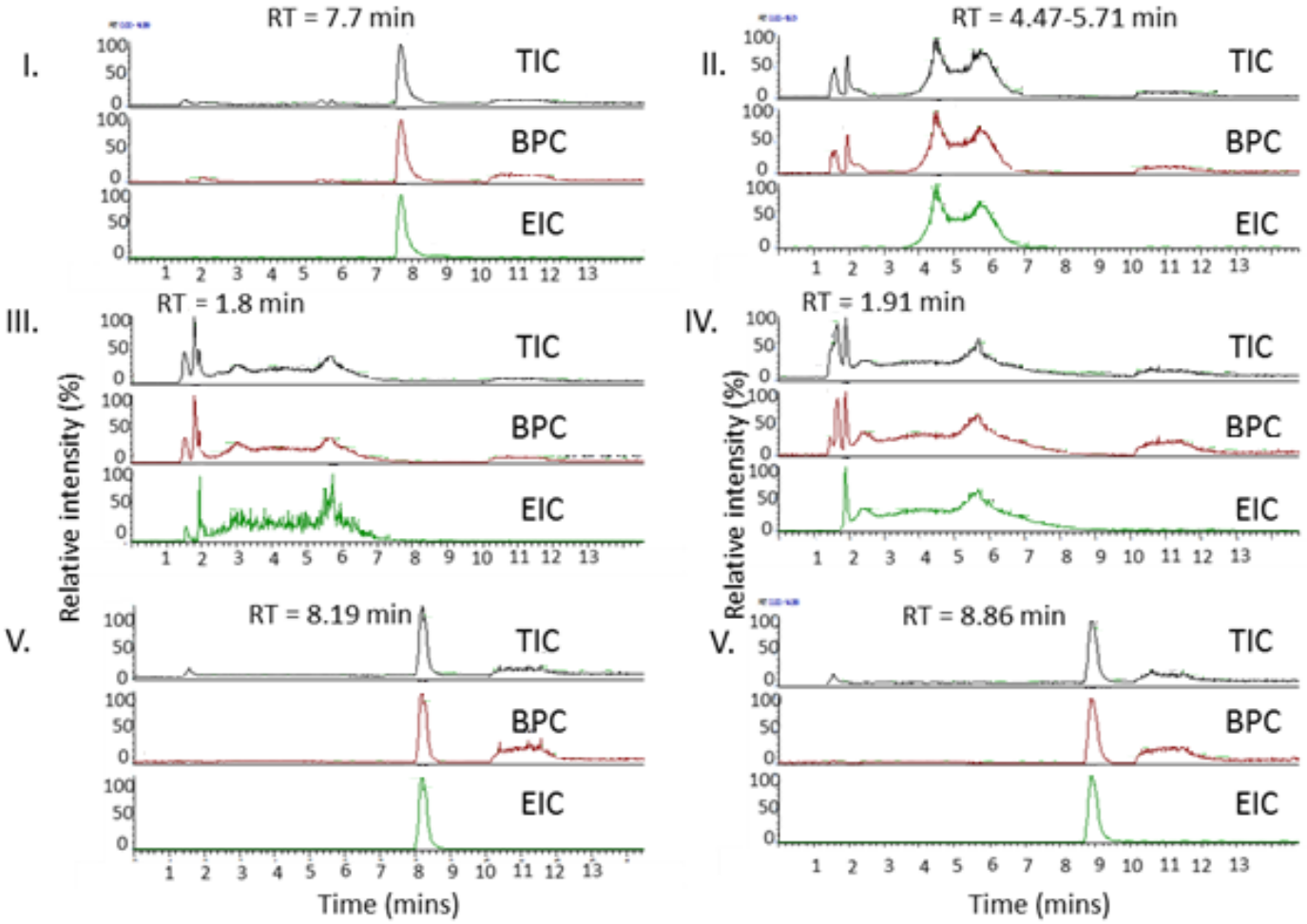


Figure 4.2.2.1: Total ion chromatogram (TIC), base peak chromtogram (BPC) and extracted ion chromatogram (EIC) generated from HILIC-full scan MS for I. IPP/DMAPP; II. GPP; III. FPP; IV. GGPP; V. DXP; and V. MEP.

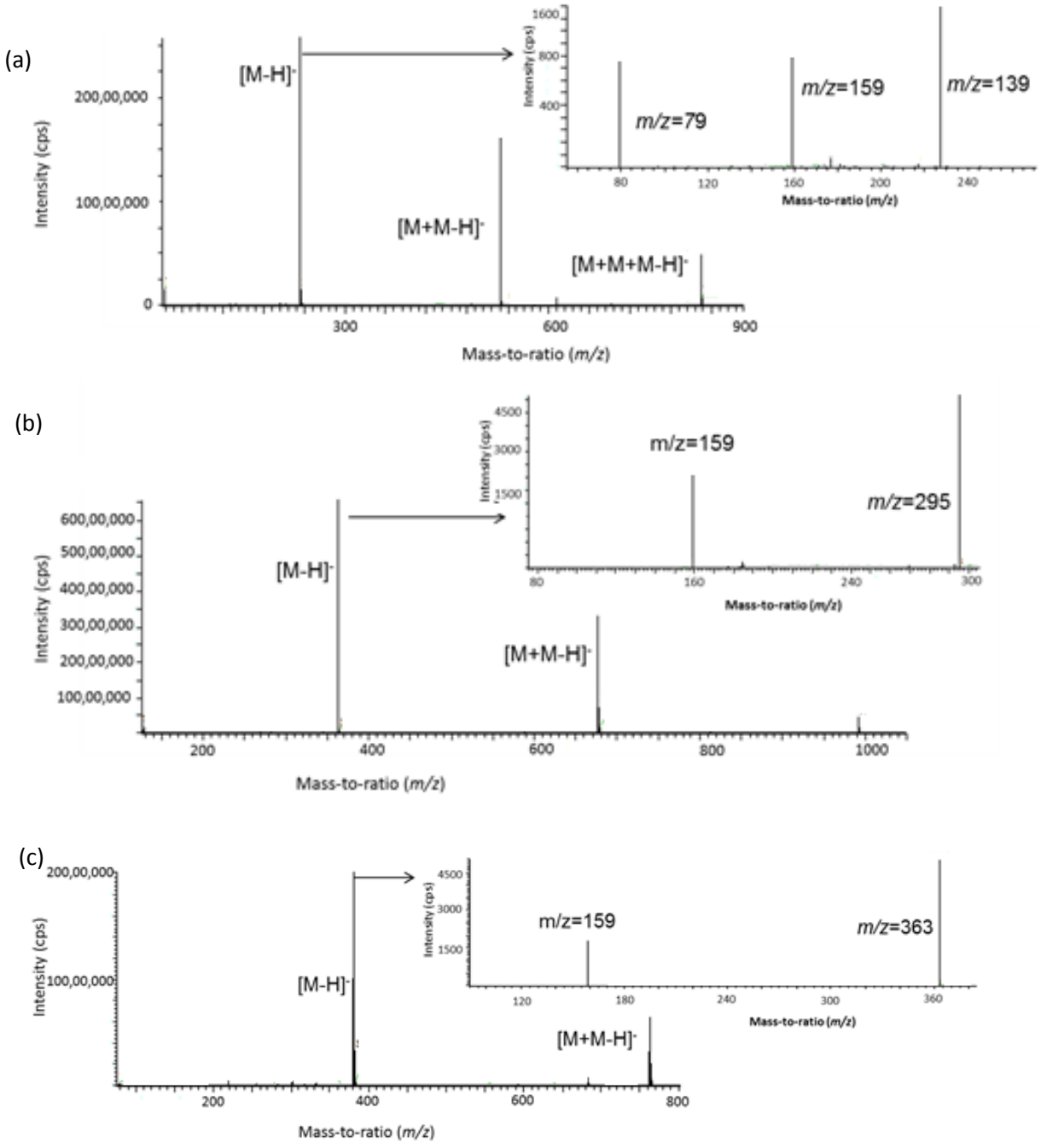
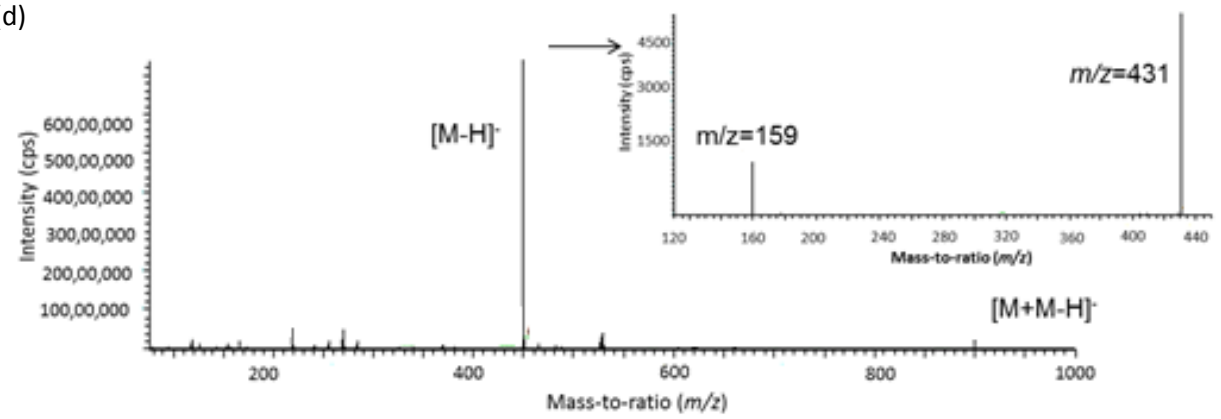
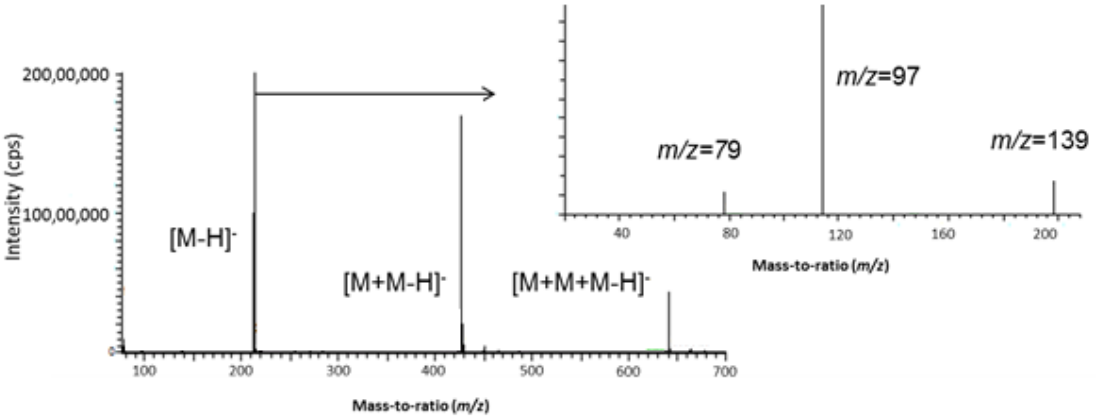


Figure 4.2.2.2: Full scan of MS for HILIC-separated isoprenoid standards for a) DMAPP/IPP; b) GPP; c) FPP. The full scan chromatograms was generated by plotting the intensity of the ions based on count per scan (cps) against the exact mass of the fragment ions. The molecular ions generated during the first scan event was monitored in the next MS/MS (inset figure).

(d)



(e)



(f)

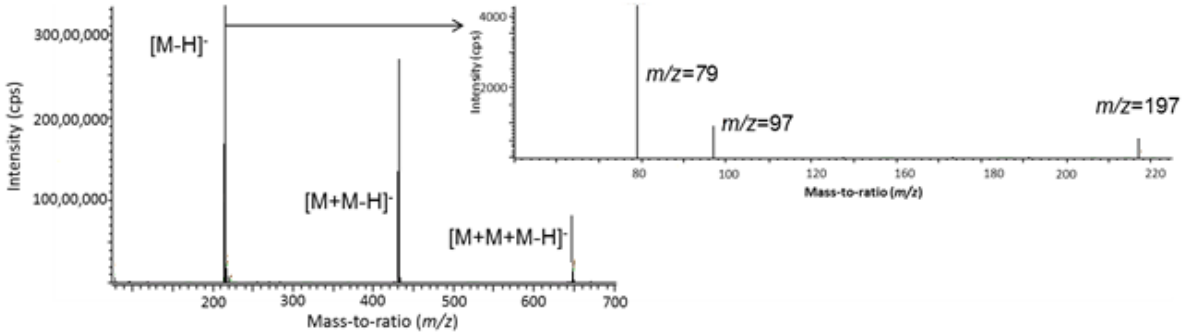


Figure 4.2.2.2 (continued): Full scan of MS for HILIC-separated isoprenoid standards for d) GGPP; e) DXP and f) MEP. The full scan chromatograms was generated by plotting the intensity of the ions based on count per scan (cps) against the exact mass of the fragment ions. The molecular ions generated during the first scan event was monitored in the next MS/MS (inset figure).

Table 4.2.2.1: List of retention time of HILIC- separated isoprenoid standards, confirmed from the extracted ion chromatogram. The electrospray ionisation (ESI)-mass-to-charge ratio (*m/z*) values was generated from the full scan MS. The collision ion dissociation (CID)-*m/z* values was obtained from data-dependent MS/MS of the molecular ion monitored from full scan MS.

Compound	**Monoisotopic mass**	**RT from MS (min)**	**ESI– *m/z* fragments**	**CID– *m/z* fragments**
IPP	246.005829	7.7	244.9988	227 159 79
DMAPP	246.005829	7.7	244.9988	227 159 79
GPP	314.068420	4.47	313.0614	295 159
GGPP	450.193634	1.91	449.1868	431 159
FPP	382.131012	1.8	381.1238	363 159
DXP	214.024246	8.19	213.0171	97 79
MEP	216.039889	8.86	215.0329	79 97 197

4.2.3. Detection of isoprenoid metabolites in plant samples

The optimised HILIC-MS/MS method was firstly utilised in identifying isoprenoid intermediates from the potato. In contrast to the samples from *Hevea* that have to be transported from its planting sites in South East Asia, potato samples could be easily accessed for method assessment. Thus, they provide readily available samples for the preliminary assessment of the developed method.

To evaluate the performance of the developed method on the potato leaf and tuber samples, the experiment was carried out as described in Chapter 2, section 2.5. Two protocols were employed to extract isoprenoids from the potato samples; protocol based on report by Li and Sharkey (2013) and a modified Folch method (Folch et al, 1951) described by Sajari et al (2014). The HILIC-MS/MS chromatograms for the potato extracts and standards are shown in Figure 4.2.3.1. Products from the two extraction methods gave rise to similar patterns of peaks, with some differences in peak signal intensities. Only peaks corresponding to DMAPP/IPP, GGPP, DXP and MEP could be identified.

The plots of peak area for DMAPP/IPP, GGPP, DXP and MEP are shown in Figure 4.2.3.2. Both extraction methods showed the GGPP compound as the highest compound detected from leaf extracts. Whilst in tuber, both methods indicated that MEP and IPP/DMAPP were predominant amongst the targeted isoprenoid compounds. The signal intensity for the HILIC-MS/MS peaks of GPP and FPP were too low and their peaks also fell outside of the expected retention times. Accurate quantification of the identified analytes cannot be performed as a standard curve was not generated from the authentic standards. Nevertheless, the level of the targeted metabolites was estimated based on the assumption of

the corresponding HILIC-MS/MS peaks fell within a linear range of quantification.

The estimation levels of the targeted compounds are detailed in Table 4.2.3.1.

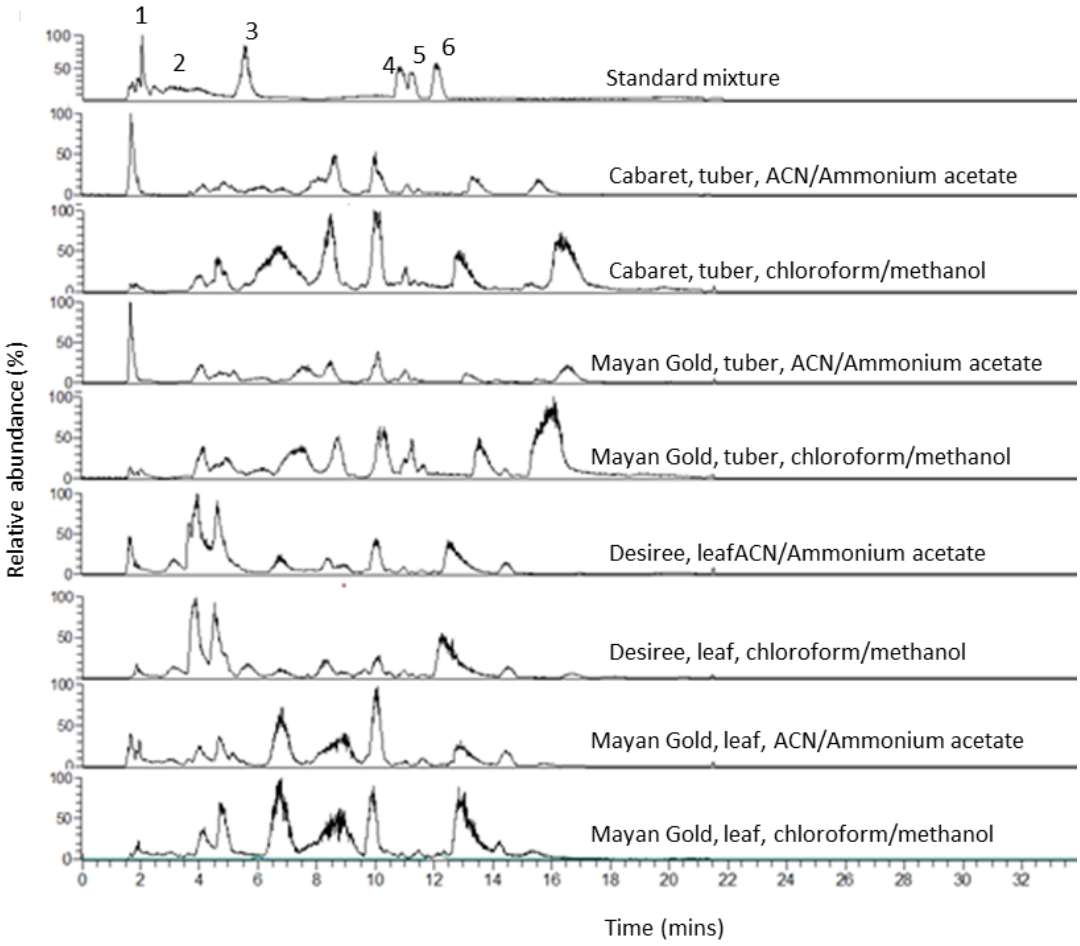


Figure 4.2.3.1: HILIC-MS/MS total ion chromatogram (TIC) for isoprenoid and potato extracts. Peaks for FPP (1), GGPP (2), GPP (3), IPP/DMAPP (4), DXP (5) and MEP (6) were identified based on the expected retention times, *m/z* values of ions generated from full MS and MS/MS scan events.

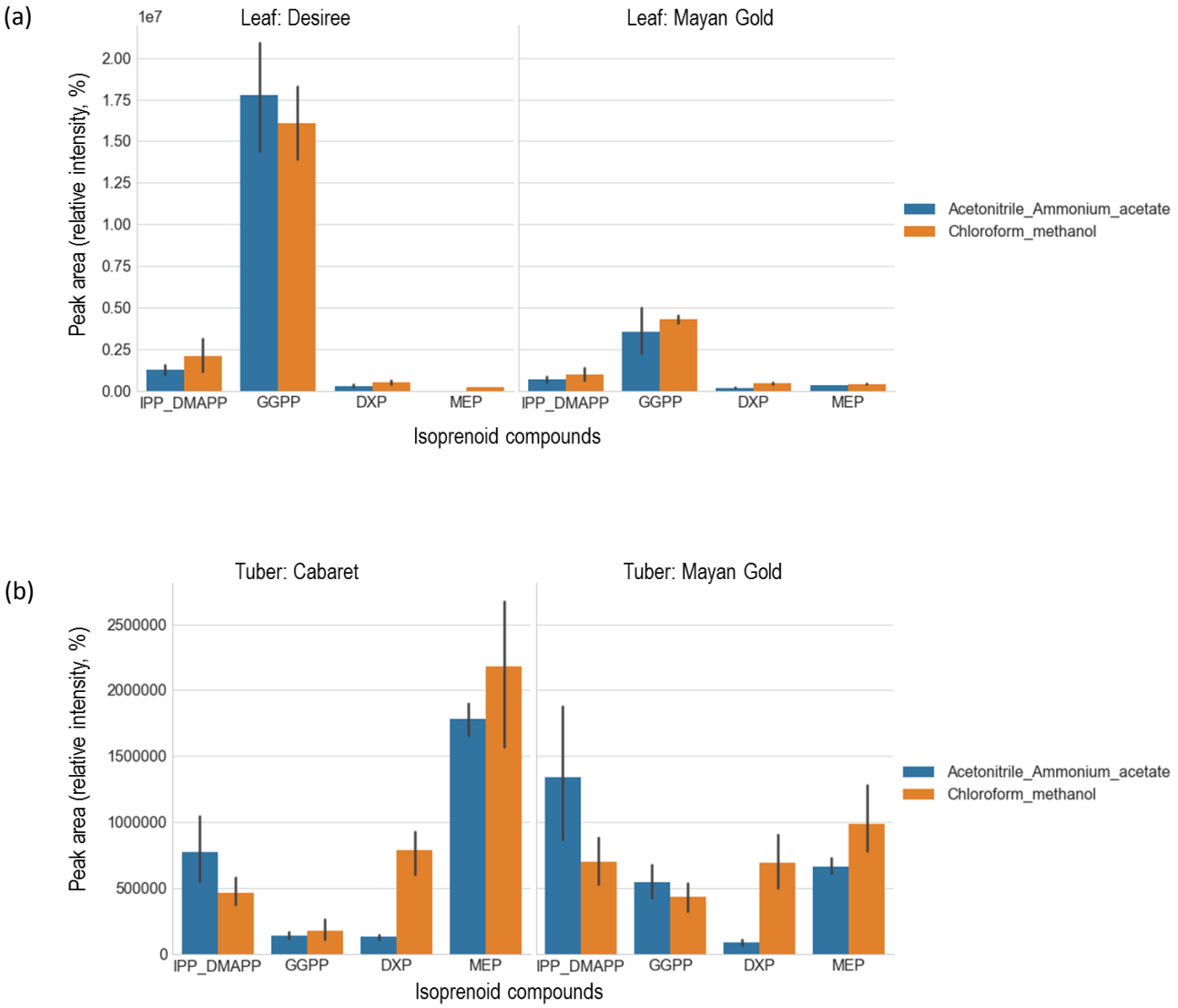


Figure 4.2.3.2: Peak areas for IPP/DMAPP, GGPP, DXP and MEP extracted using two different protocols; extraction using acetonitrile and ammonium acetate buffer (Li and Sharkey et al, 2013) and extraction using chloroform methanol (Sajari et al, 2014). Plot (a) showed extracts from leaf tissues of potato varieties Desiree and Mayan Gold. Plot (b) showed peak areas of targeted compounds from potato tuber varieties Cabaret and Mayan Gold. The whisker for each bar indicates standard deviation of the peak areas.

Table 4.2.3.1: The estimation of quantity of the identified isoprenoid compounds of different tissues of *Solanum tuberosum* and different extraction protocols.

Tissue	Variety	Compound	Extraction method	*Estimated quantity of metabolite (µg/g per dry weight)
tuber	Cabaret	IPP_DMAPP	Acetonitrile_Ammonium_acetate	0.3
			Chloroform_methanol	0.2
	Mayan Gold		Acetonitrile_Ammonium_acetate	0.6
			Chloroform_methanol	0.3
	Cabaret	GGPP	Acetonitrile_Ammonium_acetate	0.1
			Chloroform_methanol	0.1
	Mayan Gold		Acetonitrile_Ammonium_acetate	0.2
			Chloroform_methanol	0.2
	Cabaret	DXP	Acetonitrile_Ammonium_acetate	0.1
			Chloroform_methanol	0.3
	Mayan Gold		Acetonitrile_Ammonium_acetate	<0.1
			Chloroform_methanol	0.3
	Cabaret	MEP	Acetonitrile_Ammonium_acetate	0.7
			Chloroform_methanol	0.9
	Mayan Gold		Acetonitrile_Ammonium_acetate	0.3
			Chloroform_methanol	0.4
leaf	Desiree	IPP_DMAPP	Acetonitrile_Ammonium_acetate	0.5
			Chloroform_methanol	0.9
	Mayan Gold		Acetonitrile_Ammonium_acetate	0.3
			Chloroform_methanol	0.4
	Desiree	GGPP	Acetonitrile_Ammonium_acetate	7.3
			Chloroform_methanol	6.6
	Mayan Gold		Acetonitrile_Ammonium_acetate	1.5
			Chloroform_methanol	1.8
	Desiree	DXP	Acetonitrile_Ammonium_acetate	0.1
			Chloroform_methanol	0.2
	Mayan Gold		Acetonitrile_Ammonium_acetate	0.1
			Chloroform_methanol	0.2
	Desiree	MEP	Acetonitrile_Ammonium_acetate	4.3
	Mayan Gold		Chloroform_methanol	NaN
			Acetonitrile_Ammonium_acetate	NaN
			Chloroform_methanol	NaN

*The HILIC-MS/MS experiment did not possess standard curves for each of the targeted compounds. During the HILIC-MS/MS run, 20 µg individual standards were injected into the HILIC column. Therefore, the quantity estimation was made based on the comparison of peak area of the samples to that of the injected standard. The estimation was made based on the assumption that the peak of the injected standards located within the linear range of the detection.

Subsequently, the developed HILIC-MS/MS method was also used to identify the targeted isoprenoids from the latex extracts of PB235 and RRIM600 genotypes. Previous work on latex extraction optimisation (Chapter 3, carotenoid identification and quantification) showed that liquid latex resulted in a higher extraction yield compared to the coagulated latex. Therefore, in contrast to the potato sample preparation, latex extractions were performed *in-situ* in Malaysia immediately after the latex had been harvested. In addition, for the metabolite extraction, only the modified Folch method as described by Sajari et al (2014) was performed on the latex samples. This was because the modified Folch method has been successfully applied to extract metabolites from latex samples (Hasma and Subramaniam, 1986, Sajari et al., 2014) and it was found that it was easier to isolate the aqueous phase for the subsequent centrifugal-evaporation process. Two different volumes of latex (200 µl and 400 µl) were used for the extraction of latex metabolites. The dried latex extracts were transported from the extraction site in Malaysia and were reconstituted in 200 µl acetonitrile 100%:ammonium acetate 10mM pH 10 (80:20 vol/vol) prior to HILIC-MS/MS separation.

The fractionation of analytes from the 200 µl latex samples is showed in Figure 4.2.3.3. Extracts of the latex samples exhibited similar chromatogram patterns. The HILIC-MS/MS peaks were identified based on the expected *m/z* values for the targeted compounds and based on comparison of the corresponding retention time to that of its standard compound (Table 4.2.3.2). IPP/DMAPP, GPP and MEP compounds were detected from PB235 extracts whilst RRIM600 only revealed the presence of GGPP and MEP metabolites. However, due to the low signal of the peaks corresponding to the targeted

isoprenoid compounds, the MS/MS scanning did not pick up signals from their fragmented ions. Furthermore, the complexity of the MS/MS reaction was compounded by other non-targeted metabolites that were present in the latex samples at a predominantly higher level compared to that of the targeted isoprenoids. As a result, the MS/MS scanning event only detected the signals from ions with higher intensity that masked the detection of the targeted isoprenoid's ions.

Next, the HILIC-MS/MS method was applied on a higher concentration of latex extracts (from the extraction of 400 µl latex sample). However, the separation of the extracts in the HILIC column caused high back pressure that led to the termination of the HILIC-MS/MS run. Nevertheless, the truncated chromatogram profile (Figure 4.2.3.4) showed an increase in resolution of HILIC-MS/MS peaks between 3 to 12 minutes during the experiment. This indicated that higher amounts of metabolites were extracted from the 400 µl of latex sample. However, more purification steps for the latex extracts have to be performed so that impurities that caused high back pressure for the HILIC column are removed.

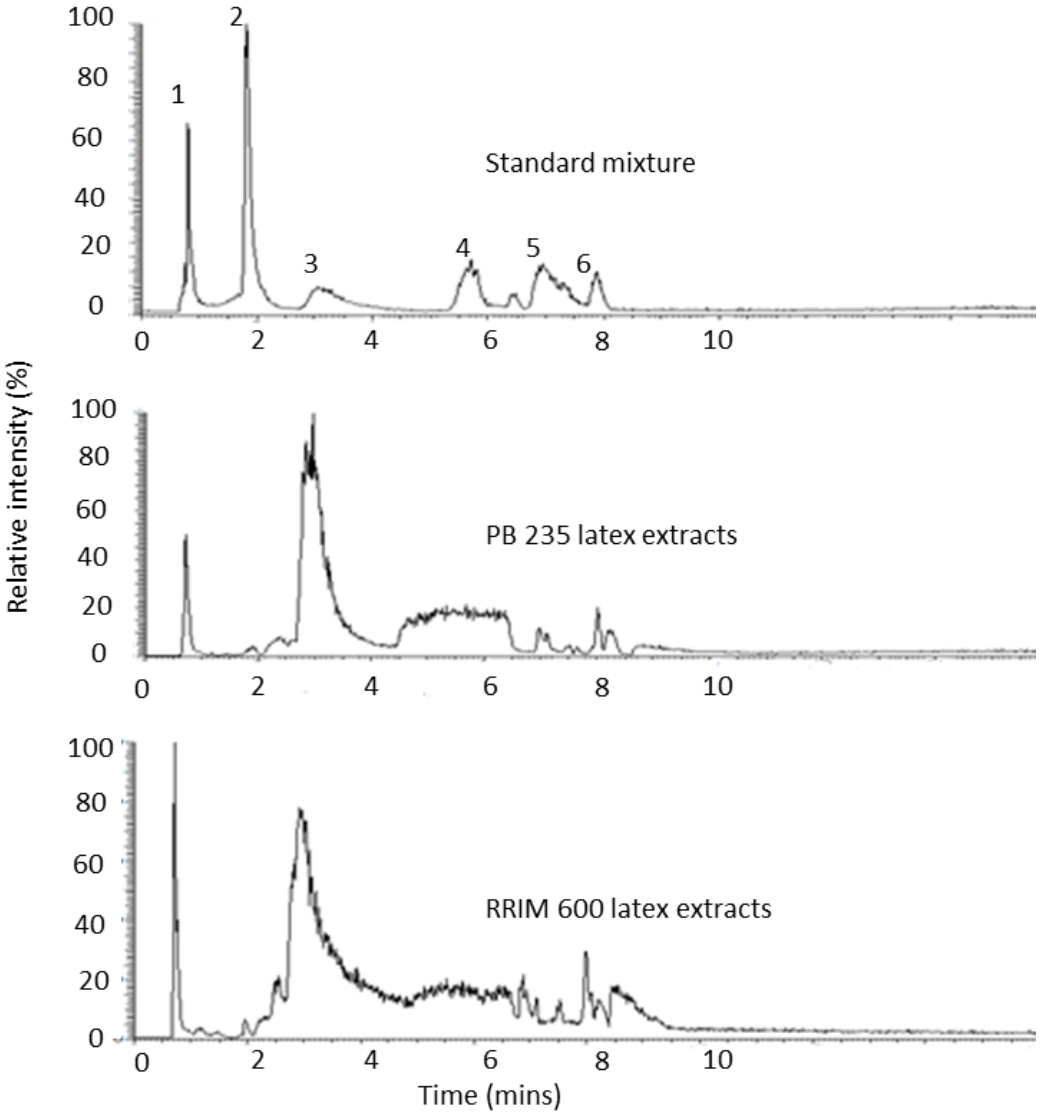


Figure 4.2.3.3: HILIC-MS/MS chromatogram of latex extracts. Peaks for FPP (1), GGPP (2), GPP (3), IPP/DMAPP (4), DXP (5) and MEP (6) were identified based on the expected retention times, *m/z* values of ions generated from full scan and MS/MS.

Table 4.2.3.2: Retention times for the targeted isoprenoid standards detected from HILIC-MS/MS chromatogram. The targeted compounds were identified from PB235 and RRIM600 latex samples by extracting peaks comparable to the retention times of the standards. The identification process was also supported by the exact parent ion mass $[M-H]^-$ detected from the full scan MS. Peak areas for the identified compounds were recorded to reflect the presence/absence of the targeted compounds from PB235 and RRIM600 latex samples. Estimation of the targeted compound was performed by comparing peak areas of the samples to that of the injected standards in the same HILIC-MS/MS run.

Targeted compounds	**Retention time (min)**	**Experimental *m/z* values**	**Peak area (relative intensity (%))**		***Estimated quantity of metabolite (µg/g per fresh weight latex)**	
			PB235	**RRIM600**	**PB235**	**RRIM600**
IPP/DMAPP	6.81	244.9989	8514.72	Not detected	0.07	Not detected
GPP	3.11	313.0613	5269.09	Not detected	0.07	Not detected
FPP	1.92	381.1238	Not detected	Not detected	Not detected	Not detected
GGPP	1.49	449.1865	Not detected	81.1154	Not detected	0.01
DXP	7.16	213.0170	Not detected	Not detected	Not detected	Not detected
MEP	7.91	215.0328	12179.47	1499.12	0.15	0.02

*The HILIC-MS/MS experiment did not possess standard curves for each of the targeted compounds. During the HILIC-MS/MS run, 20 µg individual standards were injected into the HILIC column. Therefore, the quantity estimation was made based on the comparison of peak area of the samples to that of the injected standard. The estimation was made based on the assumption that the peak of the injected standards located within the linear range of the detection.

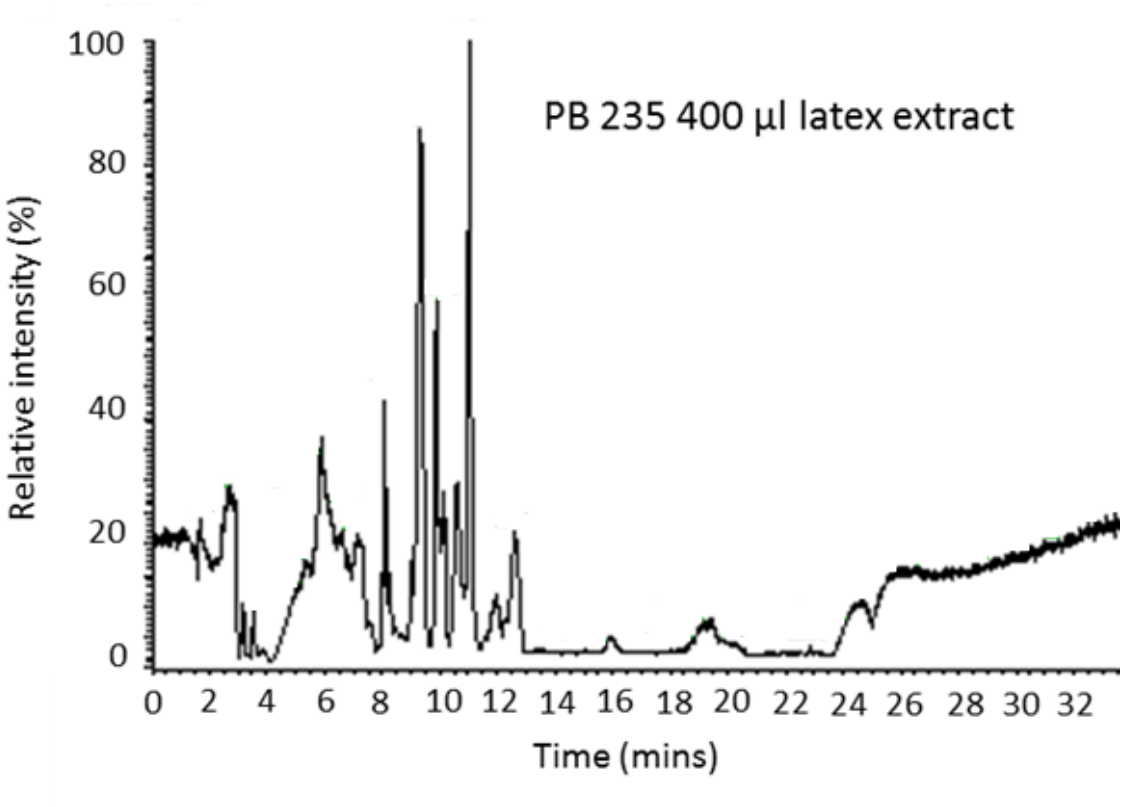


Figure 4.2.3.4: HILIC-MS/MS chromatogram for PB235 extracts, generated from 400 µl latex samples. The run was truncated at 34th minute due to the high column back pressure. Although the run was interrupted, the separation of 400 µl latex gave rise to a more diverse HILIC peaks compared to that of the of 200 µl latex extraction product.

4.3. Discussions

Metabolite profiling is one of the tools that can be utilised to give an overview of the metabolite state beyond genomic data (Bino et al., 2004). In this study, targeted metabolite profiling was applied in monitoring seven isoprenoid intermediates through HILIC-MS/MS analysis. HILIC-MS/MS was the best options to be utilised in this study as it required little volume starting materials and the HILIC-fractionated analytes could be directly linked to MS for the subsequent metabolite identification. To our best knowledge, no metabolite profiling of latex isoprenoids has been reported yet. Since there was no generic workflow specific to latex metabolite profiling that we can adhere to, the first step towards generating a robust and efficient protocol for isoprenoid profiling in latex was by qualitatively identifying the targeted compounds through accurate mass identification and its MS/MS fragments.

Satisfactory separation of the targeted isoprenoids from plant extracts, which consisted of varied polarity and molecular can be optimised by modifying the key parameters of HILIC running conditions. This included the stationary phase and the composition of mobile phase. In this study, better chromatogram peak shapes were obtained from a column with polymer-based packing materials with attached sulfobetaine group, compared to when using a column packed with silica-based with fixed amide functional group. It was reported that the polymer particles with sulfobetaine functional group provided weak electrostatic interactions of the analytes with the stationary (Buszewski and Noga, 2012). The interactions were due to the availability of both positively charged ammonium and negatively charged sulfonic within the attached functional group. This contributed to a better selectivity of the hydrophilic analytes (Guo and Gaiki, 2005). Another

important aspect contributing to the peak shape and retention of the targeted compounds during chromatography separation is the mobile phase. The HILIC method uses a mixture of ionic additives in an aqueous solvent to control the pH and ion strength of the phase. This assists the partitioning of the compound of interest either into polar or water-enriched layers that are formed within the mobile phase.

The elution order of the targeted compounds (as described in detail in the result section) was in correspondence with the molecule's polarity, followed by the molecular mass (Henneman et al., 2008, Köhling et al., 2015). The selection of the final conditions often involves a trade-off between a reasonable retention and good peak shape for all targeted metabolites and HILIC efficiency. In this study, although some of the compounds did not give a symmetrical narrow chromatogram peak, the conditions selected gave reproducible retention and reasonably acceptable peak shape for most of the metabolites in a single run. Furthermore, the tailing of GPP, FPP and GGPP peak shapes was attributed to the possibility that the isoprenoid standards might contain degradation of isoprene sub-units which may contaminate the corresponding compound. This is evidenced when mass-to-ratio (*m/z*) value of GPP could be detected in the tailing region of GGPP peak during this study.

The mass spectrometry approach was utilised in tandem to the HILIC separation to allow accurate identification of the targeted isoprenoid intermediates. As opposed to carotenoids that contain chromophore functional groups which absorb light, the targeted isoprenoid intermediates in this experiment do not absorb visible or UV light. Therefore, the targeted compounds could not be identified through the characteristic ultraviolet adsorption and instead, the peaks were identified using mass spectra. In the development of the

HILIC-MS/MS procedure in this study, the analytes separated through HILIC column were channelled into an Orbitrap Mass spectrometer. The full scan MS revealed their molecular mass (up to four decimal places) and the ions were further fragmented and analysed in MS/MS. This generated fragmentation patterns of the parent ions, in which most of the compounds showed common fragment ions at *m/z* 177 $(H_3P_20_7)^-$, 176 $(H_2P_20_7)^-$, 159 $(HP_20s)^-$, and 79 $(PO_3)^-$. These fragment ions are characteristics for diphosphate moieties, arising from the cleavage of isoprene units (Kitaoka et al., 1990). Additionally, each compound also displayed a common loss of water during fragmentation. Such fragmentation patterns (generation of diphosphate moiety and loss of water molecule) of isoprenoid compounds in MS/MS in negative mode have been reported in other system such as human tissues (Nürenberg and Volmer, 2012, Henneman et al., 2008) and plant tissues (Li and Sharkey, 2013, Wright et al., 2014). Therefore, in addition to the utilisation of retention times, metabolite verification could be further supported with the generated MS spectra comparison between the standards and the molecular ion of interest that were analysed under similar experimental conditions.

In contrast to assaying a mixture of analytical standards where a good chromatogram peak could be easily generated, analysing biological extracts using HILIC-MS/MS is more complicated. This is due to the presence of a complex mixture of biological metabolites and the acquisition of a lesser ion in MS data might be supressed by the more abundant ions. Therefore, it is necessary to evaluate the applicability of the developed method against biological extracts. In this study, most of the targeted metabolites in of the evaluated potato and *Hevea* samples did not fragment favourably and hence, no acquired MS/MS data. This is due to the more abundant ions interfering with data acquisition of

the targeted compounds present in lesser amounts. Nevertheless, the low concentration metabolites could still be identified successfully, owing to the comparison of retention times of the targeted analytes to that of corresponding authentic standards that were incorporated within the same HILIC-MS/MS run. Furthermore, by recording the experimental *m/z* values from the full scan mode within the Orbitrap mass spectrometer, the accurate masses of the targeted compounds were deciphered. The accuracy of the analytes' masses was attributed to the high sensitivity and the high mass resolving power of the Orbitrap mass spectrometer. This enabled the differentiation between non-targeted and ions of interests in a complex biological matrix as it would able to discriminate mass difference between two adjacent MS peaks (Marshall and Hendrickson, 2002).

The method was amenable to accurately quantify the targeted metabolite when standards of known amount were injected into the HILIC-MS/MS system and subjected to similar running conditions. Nevertheless, subtle discrepancies of metabolite levels could be detected between potato samples and between *Hevea* samples. In particular, the difference of MEP level could be discerned between yellow latex and white latex. Very low peak areas corresponding to IPP/DMAPP, GGPP and GGP detected from the latex extracts were not indicative of the developed HILIC-MS/MS method having less sensitivity in profiling the targeted metabolites. Rather, the low concentration of the extraction products from latex might contribute to undetected targeted compounds. However, attempts to increase the concentration of latex extracts did not give a good result, owing to the large amount of salt contaminant in the extracts. High salt contaminant in the extract caused high back pressure of the HILIC column and hence the truncated HILIC-MS/MS assay. However, an increase in the diversity

of the chromatogram peak was generated from the contaminated samples. Alternatively, the low or undetected targeted intermediates in this study could also be attributed to the concentration of isoprenoid intermediates being generally at low levels in biological tissues. This is because intermediates are turned over in the synthesis of compounds at a constant rate (Hasunuma et al., 2010).

The optimised HILIC-MS/MS can be used to simultaneously detect FPP, GGPP and intermediates from the MEP pathway (DXP and MEP compounds). Although intermediate isoprenoids were not detected (or detected at a very low level) in this study, this does not preclude the presence of isoprenoid biosynthetic genes and their corresponding enzymes in potato and *Hevea* extracts. As mentioned before, FPP and GGPP are used in the synthesis of rubber and carotenoid respectively and findings from the previous chapter (Chapter 3) showed that RRIM600 contained relatively higher amounts of rubber while PB235 latex showed relatively higher accumulation of carotenoids. Should the HILIC-MS/MS be run with increased concentration of latex extraction product, it might be possible to observe the FPP level to be higher in the RRIM600 samples compared to that of PB235. On the other hand, it is also expected to have a higher level of GGPP in PB235, rather than in the RRIM600 sample.

In conclusion, a HILIC-MS/MS technique has been developed and the optimised protocol was shown to be reproducible and could be applied to identify targeted isoprenoids from potato tissues and *Hevea* latex samples. It was necessary to generate a robust and reproducible protocol so that reliable metabolite profiling data could be undertaken. Not only is the developed method facilitate the identification of the targeted metabolites, the protocol is also amenable for future quantification of the targeted compounds.

Chapter 5

Construction of a reference transcriptome for accurate transcript profiling of the *Hevea brasiliensis* latex

5.1.1. Brief Introduction

Genomic sequences and transcriptomic resources are invaluable tools that provide information regarding the genetic potential of an organism (Spriggs et al., 2018, Abdelrahman et al., 2018, Appels et al., 2018). Unlike the model plant *Arabidopsis* that currently has the most comprehensive transcriptomic and genomics data, *Hevea brasiliensis* does not share the same breadth of resources. While genomic and transcriptomic resources for *Arabidopsis* have accumulated since the late 1990s (Rounsley et al., 1996, Lin et al., 1999, Zhu and Wang, 2000), such resources for *Hevea* only began to materialise with the first adoption of Sanger-based EST sequencing for profiling latex cDNAs (Han et al., 2013, Ko et al., 2003). Later, second generation and Sanger-based sequencing generated larger *Hevea* transcriptomic datasets, and from multiple tissues in order to understand biological processes such as rubber biosynthesis, stresses (i.e. abiotic and biotic stresses) and other aspects of biological metabolism (i.e. ethylene signaling pathway, jasmonate signaling pathway) (Chow et al., 2007, Feng et al., 2009, Montoro et al., 2008, Charbit et al., 2004). As the cost of high-throughput sequencing became cheaper and more accessible, sequencing of *Hevea* genomes and transcriptomes gained more traction (Mohamed-Sathik et al., 2018, Tan et al., 2017, Pirrello et al., 2014, Salgado et al., 2014, Wang et al., 2017, Chow et al., 2014).

Nevertheless, it still does not have the same comprehensive data provided by *Arabidopsis.* Currently, the three best available quality *Hevea* draft genomes are still not anchored to chromosomal information and contain many fragmented gene models (Tang et al., 2016, Lau et al., 2016, Pootakham et al., 2017). A

large number of scaffolds and the truncated predicted genes complicate expression measurement of target genes.

The main issue that impedes the acceleration of a high-quality assembly of the *Hevea* genome is the ambiguously mapped short reads to the repetitive genomic regions. With the repetitive regions estimated to be around 60-70% of the genome content (Tang et al, 2016, Lau et al, 2017, Poothakam et al. 2017), it is profoundly challenging only to use short reads to provide a high-quality genome assembly. Although new methods and technologies have recently become available such as long read sequencing or chromosome conformation capture (Hi-C) technique that will help in getting aa higher resolution of scaffolds, they have yet to be applied to the *Hevea* genome. Given these limitations the current project has focused on the generation of a reference transcriptome to underpin expression analysis. Such transcriptomic approach has been continually used to improve genome annotation and rare transcript identification (Tørresen et al., 2017, Sazonovs and Barrett, 2018). The utilisation of a reference transcriptome has been reported to not only improve genome annotation, but also assist in obtaining a definitive catalogue of transcript levels. The utilisation of a reference transcriptome not only useful for the non-model plant with low quality genome, it also has been applied in the plant model *Arabidopsis* for transcriptome profiling (Zhang et al, 2015, Zhang et al, 2016).

5.1.1.1 Short read sequencing

The most common protocol used to generate a reference transcriptome is through the utilisation of short read sequencing. The technology uses parallel sequencing of RNA (or cDNA) to produce a large-throughput of short

sequences (between 75 bp to 300 bp). RNA sequencing (RNA-seq) has become the standard method for transcriptomic analysis due to two main benefits: i) offering in-depth coverage of the transcriptome and ii) generating highly accurate sequenced reads. Raw RNA-seq reads now have a low error rate (0.2%), which will increase the likelihood of generating transcripts with accurate open reading frames (Manley et al., 2016, Schirmer et al., 2016).

However, the main limitation of employing RNA-seq for transcriptome profiling is in constructing correct transcript sequences from the raw reads. The construction of transcripts can be carried out *de novo* (without reference guidance) or by mapping the raw reads to the reference sequence. In this study, transcript construction was performed by assigning the reads to the draft genome sequence of *Hevea*. The main technical issue concerning the reference-based transcript construction is to unambiguously identify the position of the short RNA-seq reads in the draft genome. Inaccurate mapping of raw RNA-seq reads can lead to the generation of transcripts with unreliable exon-intron junctions. The imprecise prediction of splice-junctions in turn will result in a false representation of the transcriptome profile. Therefore, optimised parameter used during assembly/mapping is vital to ensure accuracy and completeness of the generated transcripts.

5.1.1.2 Long read technology

An alternative to short read technology is to utilise third-generation sequencing. Third-generation sequencing technology such as Pacific Biosciences (PacBio) (https://www.pacb.com/applications/rna-sequencing/) offers long-read technology, known as isoform sequencing (Iso-seq). The raw long reads can easily span the complete coding regions of transcripts. Thus,

this technology greatly facilitates the correct prediction of transcript variants without the need to assemble reads and simultaneously can also act as a framework to anchor short reads to the best position on the genome sequence.

Despite the ability to generate complete transcript sequences, the error rate in Iso-seq data is high (Weirather et al., 2017), with the primary error type being insertion/deletion (indel). If the issue is not addressed, indels will cause a shift in transcript's open reading frames and hence affect gene models. Additionally, coverage by Iso-seq is lower compared to that of RNA-seq and this leads to the lower possibility of uncovering rare transcripts.

To mitigate the effect of high error rate in the Iso-seq data, the raw reads are often corrected by either aligning the raw reads to each other or by performing hybrid sequencing. The alignment of raw Iso-seq to each other will produce consensus corrected sequences that cancel out the read errors. However, this technique of self-correction of long reads often requires high coverage. High coverage of long reads involves high sequencing cost as the throughput of the technology is not as high as that of the Illumina sequencing platform. Therefore an alternative framework involving hybrid assembly has been developed, where short reads from RNA-seq were used to correct the errors in Iso-seq (Salmela and Rivals, 2014, Hackl et al., 2014). The complementation of the short and long reads results in low-error rate, high-quality Iso-seq data.

Before Iso-seq became the most popular tool to produce long-reads, Sanger-based sequencing was considered the 'gold standard' to generate full-length cDNA sequences. This is due to the highly accurate base-calling in the Sanger sequencing technique, and this method often yields accurate full-length or near complete transcripts. However, like Iso-seq, this protocol is expensive

as it has a low-throughput sequencing rate and thus, it suffers from low-coverage of the transcriptome.

5.1.1.3 Reference transcript generation

A comprehensive transcriptome resource will provide accurate information on complete transcript sequences and the corresponding repertoire of transcript variants. To achieve a comprehensive transcriptome at a high confidence level, it is necessary to construct the transcriptome using an extensive coverage of the transcriptome dataset. A comprehensive transcriptome dataset was a prerequisite in the generation of the *Arabidopsis* reference transcript database by Zhang et al. (2017). The use of 285 raw RNA-seq datasets by Zhang et al. (2017) led to the construction of a substantial transcriptome reference that facilitated the study of alternatively spliced transcripts expressed in *Arabidopsis.* On the other hand, the available RNA-seq datasets for *Hevea* is not as comprehensive compared to that available for *Arabidopsis.* Additionally, the truncated gene models and undetected transcript variants in the *Hevea* draft genome has until now served as a as a poor reference in interpreting correct splice junctions of the constructed transcripts.

With the advent of long read sequencing technology, it is now common to use Iso-seq to improve gene model annotation and facilitate correct prediction of exon-intron junctions of transcript variants. The technique has been reported with the purpose of uncovering the complexity of transcripts in other non-model organisms such as chicken (Kuo et al., 2017), maize (Wang et al., 2016) and sorghum (Abdel-Ghany et al., 2016). Additionally, the combination of transcripts independently generated from RNA-seq and Iso-seq has been successfully used to expand the number of transcript variants

identified in barley (Mascher et al., 2017) and in mammals (Zhang et al., 2017b). Therefore, a similar framework can be applied for the generation of transcript database from *Hevea*.

5.1.1.4. Applicability of transcript database

The downstream analysis of transcriptome construction generally involves quantification of genes of interest across multiple biological conditions, exhaustive transcript annotation, identification of rare transcript variants or discrimination of allelic polymorphisms at coding region (Han et al., 2015, Hrdlickova et al., 2017). In this study, the constructed transcriptome was used to generate expression profile of key genes involved in the isoprenoid biosynthetic pathway and the transcript resource was also utilised to provide experimental evidence for the identified key genes.

The genes involved in the isoprenoid biosynthetic pathway are encoded by multi-gene families (Chappell, 1995, Hemmerlin et al., 2012). A gene family is generally defined as related genes located within a genome. The occurrence of gene families raises the question of whether each member of the gene family shares overlapping function or displays specific roles. Although many reports have been published for isoprenoid biosynthetic pathway genes in *Hevea* (Tang et al, 2016, Lau et al, 2016, Makita et al, 2017, Sando et al, 2008a, Sando et al 2008b, Chow et al 2012), the investigations still lack the information about expression for each member of the corresponding gene families. In this regard, knowledge on expression patterns of members of the gene family involved is desirable to generate a better understanding of rubber and carotenoid biosynthesis in *Hevea brasiliensis*.

5.1.2 Aims

The aim of the analysis was to produce a reference transcriptome that contains annotation of isoforms and rare transcripts beyond what is in the current *Hevea* draft genome sequence. The transcriptome was generated using raw reads obtained from multiple types of transcript data, as summarised in Figure 5.1.2.1 and in Figure S1.2.1 in Appendix. In this chapter, the merging of three independent datasets generated from RNA-seq, Iso-seq and downloaded full-length cDNA sequences is discussed (the optimisation and the construction of transcript sequences from RNA-seq data and indel correction for the Iso-seq data were described in Appendix, sections 1 - 3). Due to the large volume of RNA-seq datasets involved, the initial optimisation of the approach was performed on a small subset of the RNA-seq. The settings generated from the optimisation analysis were subsequently used for the final assembly of all RNA-seq data.

The applicability of the reference transcriptome was demonstrated through the quantification of critical genes involved in isoprenoid biosynthesis pathway. Differential expression profiles of the key genes were also obtained from two latex samples collected from two rubber tree genotypes that produced contrasting latex colours (white latex and yellow latex). As mentioned in Chapter 1, section 1.3, PB235 latex used in this study produces latex that is distinctly tinged yellow compared to the creamy white RRIM600 latex. In addition, the same samples were also applied in the metabolite profiling work (as described in Chapters 3 and 4).

The visual distinction in latex colour between PB 235 and RRIM 600 spurred the investigation of relative gene expression of selected MVA and MEP pathway steps in these genotypes by Chow et al (2012). In their study, latex

transcriptome from the latex of PB235 and RRIM600 was used to survey the isoprenoid biosynthesis pathway. However, no concurrent analyses of the expression level of genes involved in isoprenoid biosynthesis and levels of isoprenoid-related metabolites in latex has been carried out to investigate the relationship between pathway gene expression and end product level. This chapter explores the needs to generate a reference transcriptome and describes the utility of the reference transcriptome in creating expression patterns of genes in the latex of PB235 and RRIM600 rubber tree genotypes. The downstream analysis shows the level of expression of the essential genes and the preferential expression of specific isoforms within latex.

5.2. Results

5.2.1. Survey of the *Hevea* draft genome

The survey of the *Hevea* genome assemblies was performed to investigate the *Hevea* draft genome quality. A finished or high-quality genome sequence is defined as one that contains no or minimal gaps, high contiguity, with high accuracy at the base level, gene level and pseudomolecule level (Mardis et al., 2002, Chain et al., 2009, Sedlazeck et al., 2018). However, attaining high-quality genomic data requires laborious and cost-extensive work (with regard to the library preparation and data processing). Achieving a high-quality genome also depends on the genome organisation of the organism of interest. For example, an organism with a highly heterozygous genome will pose a high risk of mis-assembly of the genome data (Kajitani et al., 2014, Shimizu et al., 2017). Additionally, a genome with a high proportion of repetitive

region commonly gives a large number of unassembled regions or gaps (Bevan et al., 2017, Hulse-Kemp et al., 2018). Generally, during genome assembly, scaffolds are assembled from adjacent contigs (a subset of fragments of genomic DNA sequence generated from aligned raw reads). During the scafolding process, contigs are chained in the correct order and orientation by gaps (strings of 'N' letters). If important gene features fall within such gap regions, they will not be detected.

Due to these technical issues, most of the non-model genome resources are often presented at the unfinished stage (also known as a draft genome). Nevertheless, the draft genome is still useful for subsequent analysis (for example, sequence searches of a gene of interest) as that is the most current genomic resource available. However, before the downstream analysis commences, it is essential to sample the quality of the information available in the draft genome. This would give an overview of the limitation of the genomic resources and helps in determining the most suitable approach for subsequent analysis. In *Hevea,* four versions of draft genomes were reported from four rubber tree genotypes, namely RRIM928 (Mollison et al., 2014), RRIM600 (Lau et al., 2016), Reyan 7-33-97 (Tang et al., 2016) and BPM24 (Pootakham et al., 2017). Table 5.2.1.1 lists the basic statistics of the four available draft genomes. It shows that RRIM928 has the most truncated assemblies as it contains the highest number of scaffolds and yet, the lowest value of its N50. In contrast, the RRIM600 assembly showed the highest coverage of *Hevea* genome, its corresponding number of scaffolds is higher compared to that of Reyan 7-33-97. Although the BPM24 assembly shows a comparable number of predicted genes and length of assembled genome to that of the Reyan 7-33-97 draft genome, the number of scaffolds is larger. This indicates that Reyan 7-33-

97 draft genome has the highest contiguity of assembly. Therefore, based on the contiguity of assembly, manual curation of selected gene models were performed using predicted genes from the Reyan 7-33-97 draft genome.

An overview of gene model completeness for each was evaluated by manual inspection of gene models of genes from the isoprenoid biosynthesis pathway, namely lycopene β-cyclase (LCY), acetyl-coA C-acetyltransferase (AACT), 1-Deoxy-D-xylulose 5-phosphate synthase (DXS) and geranyl diphosphate synthase (GPPS). The gene models were searched using TBLASTN analysis as described in Chapter 2, (section 2.29). Manual inspection of the gene models showed two main issues; (i) truncated gene model due to mis-assembly and due to scaffolding; and (ii) minimal information regarding alternatively spliced products for each gene model. For example, a coding region that is punctuated by scaffolding will have interrupted open reading frames and this may impact downstream analysis related to gene regulation and functions. The example of observations of the gene model quality is summarised in Figure 5.2.1.1. The effect of using a truncated draft genome will greatly impact the accuracy of the downstream analysis. Despite having four versions of the *Hevea* draft genome, an accurately annotated genome sequence is not available. Thus an approach to quantify gene expression based on the generation of a reference transcriptome appears to be the best choice for the present study.

Table 5.2.1.1 Basic statistics of *Hevea* draft genomes.

Genotype	Number of scaffolds	N50 (kb) of contigs	N50 (kb) of scaffolds	Number of genes	Estimated genome size (Gb)	Assembled genome size (Gb)
Reyan 7-33-97	7,453	30.6	1,280.00	43,792	1.46	1.37
RRIM600	189,316	20.8	67.24	84,440	2.15	1.55
BPM24	592,580	2.9	96.83	45,236	Not indicated	1.26
RRIM928	954,373	Not indicated	26.81	Not indicated	Not indicated	1.24

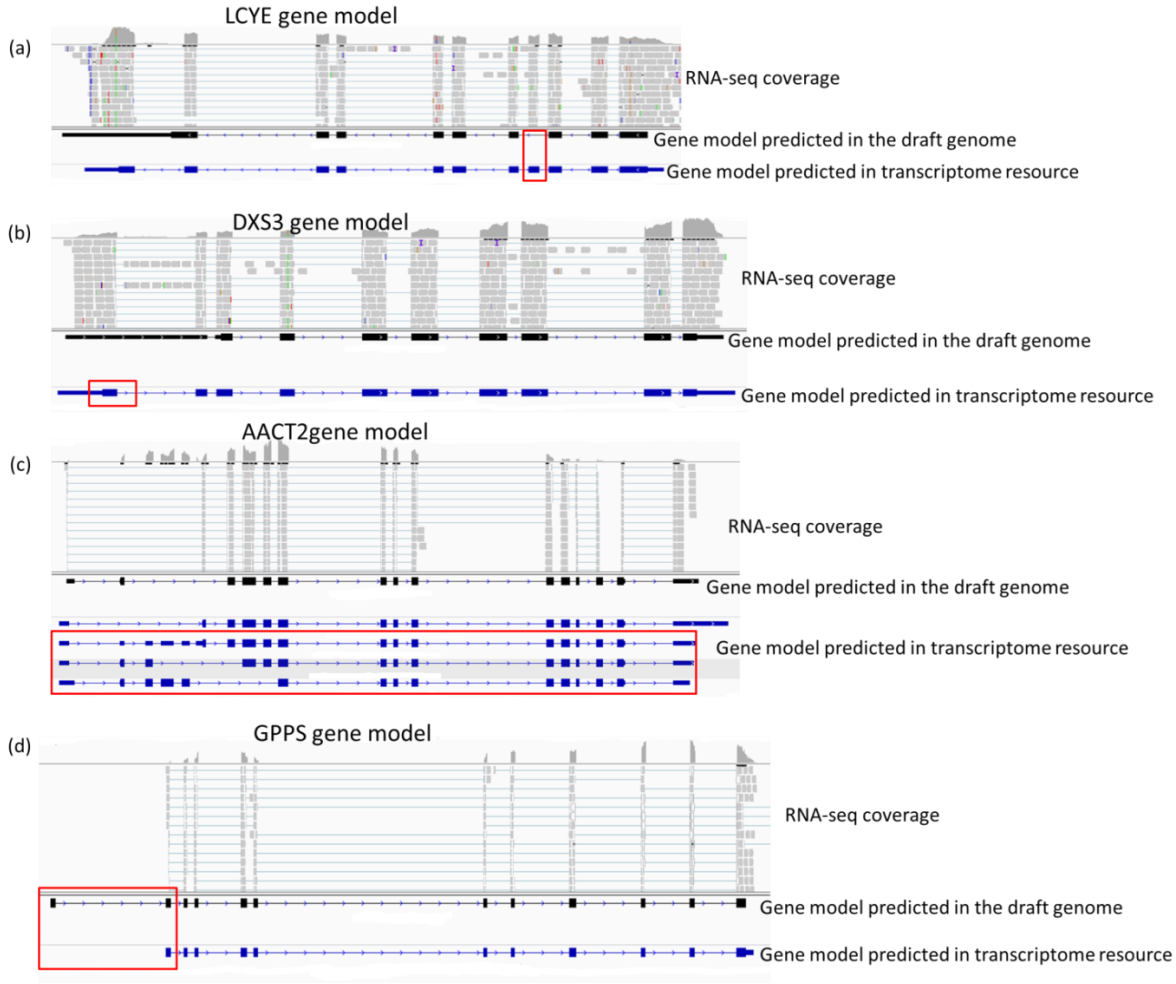


Figure 5.2.1.1: Examples of sequence inaccuracies in the predicted gene models. The predicted gene models were truncated in LCYE gene model (a) and in DXS3 gene model (b). The truncated regions are marked in red rectangles. The AACT2 gene model (c) was updated by different alternatively spliced products predicted from the transcriptome resource (marked in the red rectangle). Finally, GPPS gene model (d) was found to be mis-assembled as RNA-seq coverage was not found to support the first exon that was predicted in the draft genome sequence. The unsupported region is marked in a red rectangle.

5.2.2. Construction and evaluation of a reference transcriptome

Transcript profiling can be carried out using a reference transcriptome that ideally encompasses comprehensive alternatively spliced products and accurate open reading frames of all transcripts. The construction of the current reference transcriptome of *Hevea* involved three types of data namely RNA-seq, Iso-seq and full-length cDNA sequences.

Firstly, RNA-seq data from Illumina's Next Generation Sequencing platform were used in map-based transcript construction. Although the read length was short (150 bp), its high throughput and low error rate during base-calling facilitates transcript coverage and in uncovering rare transcripts that were not annotated in the draft genome sequence. Secondly, Iso-seq sequences generated by Pootakham et al (2017) were also incorporated in the construction of the reference transcriptome. The salient feature of Iso-seq data is that the long sequence technology generates a complete sequence of transcript variants. With this capacity, an accurate coding region of isogenes (or gene members) and transcript variants can be produced simultaneously. Thirdly, highly redundant full-length cDNA sequences generated by Makita et al (2017) were also used in the transcript reference catalogue construction. Like Iso-seq data, the full-length cDNAs provide *bona fide* transcript structure, due to the long-length and the highly accurate base-calling of Sanger sequencing technology (Heather and Chain, 2016).

By merging independent gene evidence from RNA-seq, Iso-seq and full-length cDNAs, an in-depth, high coverage and accurate reference transcriptome was generated (for at least the latex tissue). The reference transcriptome construction in this study was carried out according to a method described in

Chapter 2 (sections 2.21-2.28). Afterwards, evaluation of the merged transcripts was performed using protocols as described in Chapter 2 (section 2.29). For merging of the reference transcriptome, 76,596 transcripts from RNA-seq data, 194,135 error-corrected Iso-seq sequences and 24,327 full-length cDNA sequences (Makita et al., 2017) were merged into a single reference transcriptome. The construction of reference-based transcripts from RNA-seq and processing of the Iso-seq data are described in Appendix A sections 1-3. The redundant sequences in the transcript catalogue were reduced by software from the COGENT pipeline (Tseng, 2016). The COGENT pipeline was selected as it enabled the identification of highly identical gene members, without depending on mapping to a reference genome.

Table 5.2.2.1 summarises (a) the clustering results of the merged transcripts and (b) basic sequence statistics of the merged transcripts. A total of 182,998 transcripts were collapsed into 24,704 gene families. The remaining 112,060 sequences that were not grouped into gene families were classified as orphans (or putative single-member genes). COGENT has clustered 193,997 sequences as a final set of transcripts (with reduced redundancy), with an average sequence length of 1,671 bp. Based on BLASTN analysis, more than 91% of the merged transcripts can be matched to that of the transcripts annotated in the *Hevea* draft genome (Tang et al, 2016). Additionally, 28,301 of the merged transcripts were found to correspond to transcripts annotated from the *Hevea* draft genome. BUSCO analysis (Figure 5.2.2.2) indicated that the merged transcripts contained the highest coverage of plant's essential single-copy orthologs.

Table 5.2.2.1: Details of (a) the clustering results of the merged transcripts and (b) basic sequence statistics of the merged transcripts

(a) Cogent Clustering

Clustering results	Counts
Merged transcripts (prior to clustering)	295,058
Total grouped by COGENT	182,998
Orphan sequences	112,060
Merged, transcripts	193,997
Gene families number predicted by COGENT	24,704

(b) Sequence statistics of the merged transcriptome after the redundancy reduction.

Statistic	Counts
Average length (bp)	1,671
Longest transcript (bp)	28,240
N50 (bp)	2,102
Transcripts with 1:1 relationship[#]	28,301
Percentage of the merged transcripts matched* to the annotated transcripts (Tang et al, 2016)	91% (177,001 of the merged transcripts matched to the annotated transcripts)

[#] Obtained from reciprocal best BLASTN hit between the transcripts and annotated transcripts from the *Hevea* draft genome (Tang et al., 2016).
* BLASTN hits were filtered based on >95% percentage identity, >50% query coverage.

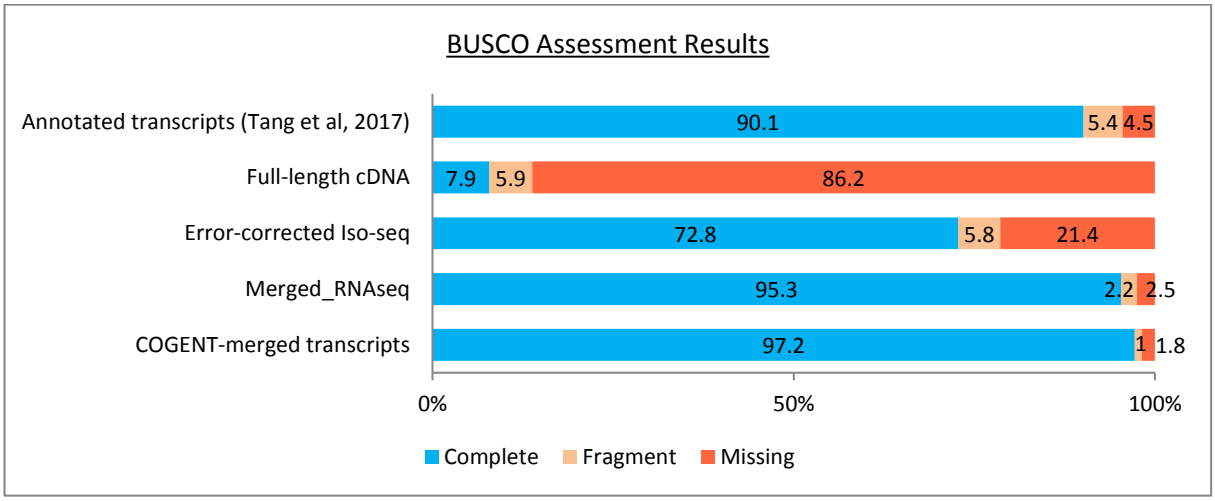


Figure 5.2.2.2: Assessment of the completeness of the transcriptome using BUSCO. The completeness was represented by BUSCO through the coverage of single copy orthologs that are essential and should be present across plant kingdom. The merged transcript constructed in this study has shown the highest coverage of the single copy orthologs compared to other transcriptome data.

Despite using the COGENT pipeline to merge the transcripts based on their sequence similarity, some of the gene families still exhibited redundant transcripts. The redundancy was typically marked by multiple transcripts bearing identical coding region and yet varied lengths of the untranslated region (UTR), with one example is depicted in Figure 5.2.2.3. Though these may reflect differences in either transcription start sites or polyadenylation position, due to the method of transcript capture and sequencing (used to garner the transcript data) this ambiguity could not be resolved.

To address this issue, manual curation of the reference transcriptome was performed on the genes involved rubber and carotenoid biosynthesis pathways. After manual inspection, only transcripts representing the longest UTR was retained. Additionally, redundant transcripts encoding truncated gene models were discarded from the set of curated transcript dataset. The manual curation must be performed due to the time limitation that hampered the effort of developing an in-house script for the efficient removal of redundant transcripts from the reference transcriptome. The curated transcripts are listed in Table S3.2 in Appendix A.

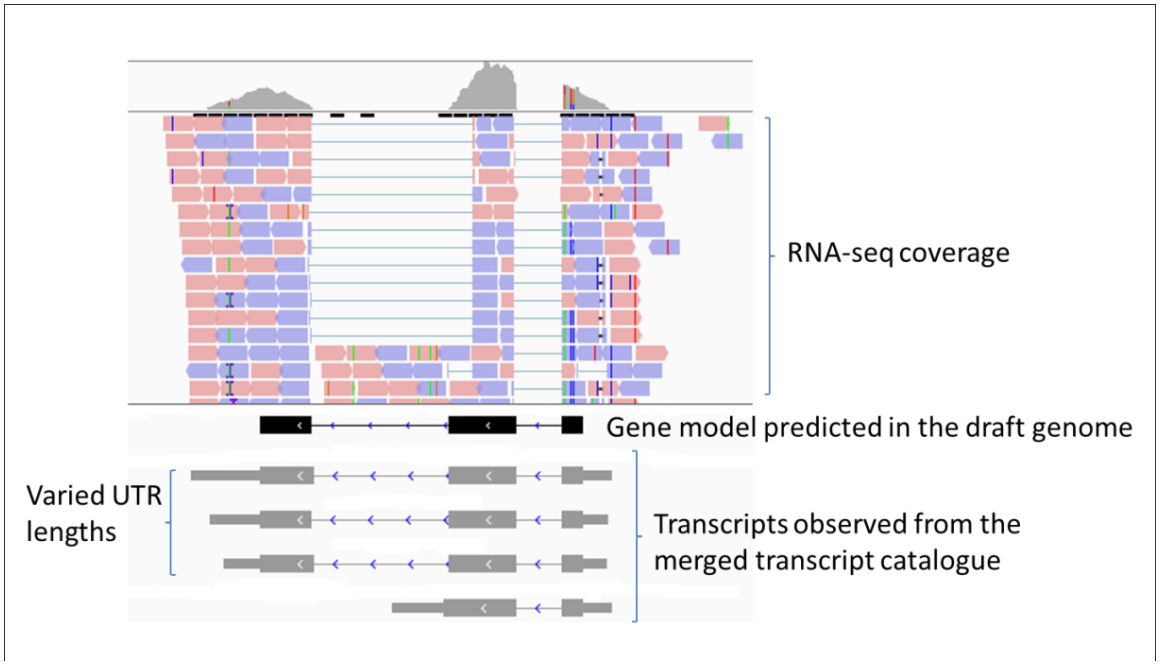


Figure 5.2.2.3: An example of a gene model with its associated predicted transcripts, viewed using Integrative Genomics Viewer (IGV) software (Thorvaldsdóttir et al., 2013). The gene model was predicted by Tang et al (2016) in the *Hevea* draft genome. The associated transcripts were obtained from the merged transcripts catalogue. Even though COGENT would reduce the redundancy in the merged transcripts, some gene models were still observed to have identical transcription start and termination sites. The only difference that these transcripts have was that they have different length of 3' untranslated region (UTR).

5.2.3. Identification of key genes involved in carotenoid and rubber formations

The key enzymes involved in carotenoid and rubber formations in *Hevea* latex were described in Chapter 1, section 1.3 and Chapter 3, section 3.1.1. These enzymes have been reported to play important roles on the accumulation of these two isoprenoid metabolites in *Hevea* latex and other plant species. However, the transcript information on which it was based was not comprehensive and is likely contain inaccurate evidence based on coding region fragment and limited transcript variants. Thus, in the current study, a manually-inspected reference transcriptome based on the key pathways and genes of interest has been used in the transcriptome profiling of the latex of PB235 and RRIM600. A total of 115 gene models on 97 scaffolds of *Hevea* draft genome, with 151 corresponding transcript variants of the target genes have been compiled on which to base the expression profiling analysis (Table 5.2.3.1). These 151 transcripts can be classified into 115 enzyme categories, associated with the five metabolic pathways that ultimately produce rubber and carotenoid. The five metabolic pathways were as follows: 1) MVA pathway, 2) MEP pathway, 3) isoprenoid initiator formation, 4) rubber elongation steps and 5) the carotenoid biosynthetic pathway. Of the 115 gene models, 19 were identified for MVA pathway, 22 for the MEP pathway, 16 isoforms encoded genes for the formation of isoprenoid initiators, 31 isoforms for rubber elongation steps and 27 isoforms belonged to the carotenoid biosynthetic pathway.

In this study, gene members (or isoforms) are defined as sets of highly similar proteins (or DNA coding regions) which each originates from a unique gene model. This includes duplicated gene models that are located on a

different genomic location. For example, AACT gene family predicted from the draft genome constituted of two genes (*AACT1* and *AACT2*). Each gene is duplicated on two different scaffolds (*AACT1* on scaffold0792 and scaffold0001; *AACT2* on scaffold1479 and scaffold0992). Therefore, in this study, AACT gene family is encoded by four different gene members, based on the four gene models identified from different genomic locations. However, it should be noted in a highly heterozygous draft genome, it is possible that at least some gene members may represent different alleles at an individual locus. This may have arisen during assembly, where they have been separated into different contigs or scaffolds rather than collapsing into a single representative sequence.

Table 5.2.3.1: Number of genes of interest, number of isoforms, number of alternatively spliced products. The transcriptome resource has predicted a larger number of alternatively spliced products compared to that of the draft genome.

Pathway	Gene of interest	Scaffold	Number of identified isoforms
MVA	Acetyl-CoA C-acetyltransferase (AACT)	scaffold0792, scaffold0001, scaffold1479, scaffold0992	4
	Hydroxymethylglutaryl-coa synthase (HMGS)	scaffold1236, scaffold0082	2
	Hydroxymethylglutaryl-coa reductase (HMGR)	scaffold0419, scaffold0592, scaffold0162, scaffold1265, scaffold0896, scaffold0344	6
	Mevalonate kinase (MK)	scaffold0374, scaffold0013	3
	Phosphomevalonate kinase (PMK)	scaffold1109, scaffold1148	2
	Mevalonate diphosphate decarboxylase (MVD)	scaffold0676, scaffold0055	2
	Total		19
MEP	1-Deoxy-D-xylulose 5-phosphate synthase (DXS)	scaffold0007, scaffold0155, scaffold0816, scaffold0094, scaffold0615, scaffold0582, scaffold0458, scaffold0067, scaffold0795, scaffold0610	10
	1-Deoxy-D-xylulose 5-phosphate reductoisomerase (DXR)	scaffold0121, scaffold1113	2
	2- C-methyl-D-erythritol 4-phosphate cytidylyltransferase (MEPCT)	scaffold0818, scaffold0300	2
	4-(Cytidine 5-diphospho)-2- C-methyl-D-erythritol kinase (CDPMEK)	scaffold0014, scaffold0056	2
	Cyclodiphosphate synthase (MECPS)	scaffold0103, scaffold0149	2

	4-Hydroxy-3-methylbut-2-enyl-diphosphate synthase (HDS)	scaffold1509, scaffold0574	2
	4-Hydroxy-3-methylbut-2-enyl diphosphate reductase (HDR)	scaffold0731, scaffold0309	2
Total			22
Formation of isoprenoid initiators	Isopentenyl diphosphate Δ-isomerase (IPPI)	scaffold0868, scaffold0177	2
	Geranyl diphosphate synthase (GPPS)	scaffold1292, scaffold0221, scaffold0066, scaffold0552	5
	Farnesyl diphosphate synthase (FPPS)	scaffold0157, scaffold0411, scaffold0908, scaffold0916	4
	Geranylgeranyl diphosphate synthase (GGPPS)	scaffold0476, scaffold0724, scaffold0434, scaffold0131, scaffold1429	5
Rubber elongation steps	Rubber elongation factor (REF)	scaffold1222, scaffold0818	8
	Small rubber particle protein (SRPP)	scaffold1222, scaffold2538, scaffold0197, scaffold0824, scaffold0916, scaffold0624	10
	cis-prenyl transferase (CPT)	scaffold0385, scaffold0387, scaffold1517, scaffold2210, scaffold2806	6
	Rubber biosynthesis stimulator (RBSP)	scaffold0224, scaffold0390, scaffold0726, scaffold0872, scaffold2400	5
	Rubber biosynthesis inhibitor protein (RBIP)	scaffold0100	2
	Total		47
Carotenoid	Phytoene synthase (PSY)	scaffold0036, scaffold0786, scaffold0440, scaffold0494	4
	Phytoene desaturase (PDS)	scaffold0026	1
	Zeta carotene desaturase (ZDS)	scaffold0239, scaffold0114	2
	15-cis-ζ-carotene isomerase (Z-ISO)	scaffold0587	1
	Carotenoid isomerase (CRT-ISO)	scaffold0099, scaffold2091	2

	Lycopene β-cyclase (LCYB)	scaffold0303	1
	Lycopene epsilon-cyclase (LCYE)	scaffold0642	1
	β-carotene hydroxylase (BCH)	scaffold0143, scaffold1120, scaffold0103	3
	Zeaxanthin epoxidase (ZEP)	scaffold0245, scaffold0420	2
	Violaxanthin de-epoxidase (VDE)	scaffold1069	1
	Neoxanthin synthase (NSY)	scaffold0814	1
	9-cis-epoxycarotenoid dioxygenase (NCED)	scaffold0866, scaffold1099, scaffold1016, scaffold0035, scaffold0641, scaffold0506, scaffold0085, scaffold2361	8
	Total		27

5.2.4. Utilisation of the reference transcriptome in the analysis of differential expression of isoprenoid biosynthetic genes

The expression analysis was performed on transcripts which play a major role in the accumulation of carotenoids and rubber in the latex of *Hevea.* Such expression levels studies have been reported previously by several groups, by counting RNA-seq reads that were either inferred from the draft genomic sequence (Hurtado-Páez et al., 2015) or based on a set of reference transcripts or by *de novo* set of reference transcripts (Xia et al., 2011, Salgado et al., 2014, Mohamed-Sathik et al., 2018). It has been established that using such reference sequences (either an unfinished draft genome or *de novo* transcriptome) would give ambiguous information regarding the expression profile for the truncated gene models or redundant transcripts (Brown et al., 2017b, Zhang et al., 2016). In this study, the utility of the constructed reference transcriptome was demonstrated through the differential expression of the targeted genes (identified in section 5.2.3) from PB235 and RRIM600 latex samples.

The analysis of differential expression of the targeted genes based on the manually curated transcriptome was performed by firstly constructing the RNA-seq data from the latex of the rubber tree genotypes PB235 and RRIM600. A total of six library replicates were constructed for each rubber tree genotype. The preparation of the libraries was carried out according to the protocols as described in Chapter 2 (sections 2.3.1 – 2.3.6). In addition, to get an overview on the expression of targeted key genes involved in the isoprenoid biosynthetic pathway within multiple tissues, RNA-seq data generated from the rubber tree genotype Reyan 7-33-97 (downloaded from the NCBI database) was also used. Although the downloaded RNA-seq datasets did not have

replication in contrast to the libraries constructed in this study, they covered a wide range of tissue and this would give at least an indication about tissue level specificity. Secondly, the transcript profiling from RNA-seq reads was performed by counting the reads that mapped to the manually curated transcriptome. The differential expression profiles were inferred based on the values of the transcript per million (TPM), which was calculated based on the number of mapped reads in the normalised RNA-seq data. The expression profile analysis was performed according to steps as mentioned in Chapter 2 (sections 2.3.8.1 and 2.3.13). A total of 467,118,723 of cleaned read pairs (ranging from 14,240,119 to 55,915,054 for each RNA-seq of PB235 and RRIM600 constructed libraries) were utilised for the comparison of expression profile between PB235 and RRIM600 (Table 5.2.4.1). For an overview of expression of targeted genes across multiple tissues of Reyan 7-33-97, 195,006,842 downloaded reads were cleaned and used for the expression measurement.

The normalisation or down-sampling of the mapped reads (to ensure that the read depth between samples was equivalent) was done using EdgeR software (Robinson et al., 2010). The normalisation step was necessary as it adjusted read count based on the normalisation factor and hence corrected for differences in total read count between samples. The multidimensional scaling plot (MDS) for each dataset was generated using Limma software (Law et al., 2016) and gave an overview of the variations between the normalised expression data. Figure 5.2.4.1 illustrates the MDS plot of PB235 and RRIM600. The MDS indicated the six RRIM600 datasets clustered into the same group. However, it was observed that while four of the PB235 datasets clustered into a single group, two of the datasets that were supposed to belong to PB235 (PB235.3 and PB235.4) were found to cluster closer to RRIM600.

Table 5.2.4.1: RNA-seq datasets used for transcript profiling analysis and the corresponding total number of cleaned reads.

Latex colour/tissue	**RNA-seq library/rubber tree genotype**	**Cleaned paired reads**
Yellow latex (from PB235 rubber tree genotype)	PB235.3	33,736,504
	PB235.4	14,240,119
	PB235.5	38,817,516
	PB235.6	18,553,640
	PB235.7	55,915,054
	PB235.8	41,156,865
White latex (from RRIM600 rubber tree genotype)	RRIM600.3	41,171,412
	RRIM600.4	47,494,075
	RRIM600.5	53,702,318
	RRIM600.6	36,539,191
	RRIM600.8	33,107,909
	RRIM600.9	52,684,120
Latex (NCBI accession number: SRR3136162)	Reyan 7-33-97	13,224,025
Bark (NCBI accession number: SRR3136158)		18,019,834
Leaf (NCBI accession number: SRR3136159)		20,218,651
Female flower (NCBI accession number: SRR3136165)		29,411,179
Male flower (NCBI accession number: SRR3136166)		26,242,643
Seed (NCBI accession number: SRR3136168)		24,557,762
Root (NCBI accession number: SRR3136156)		25,853,282

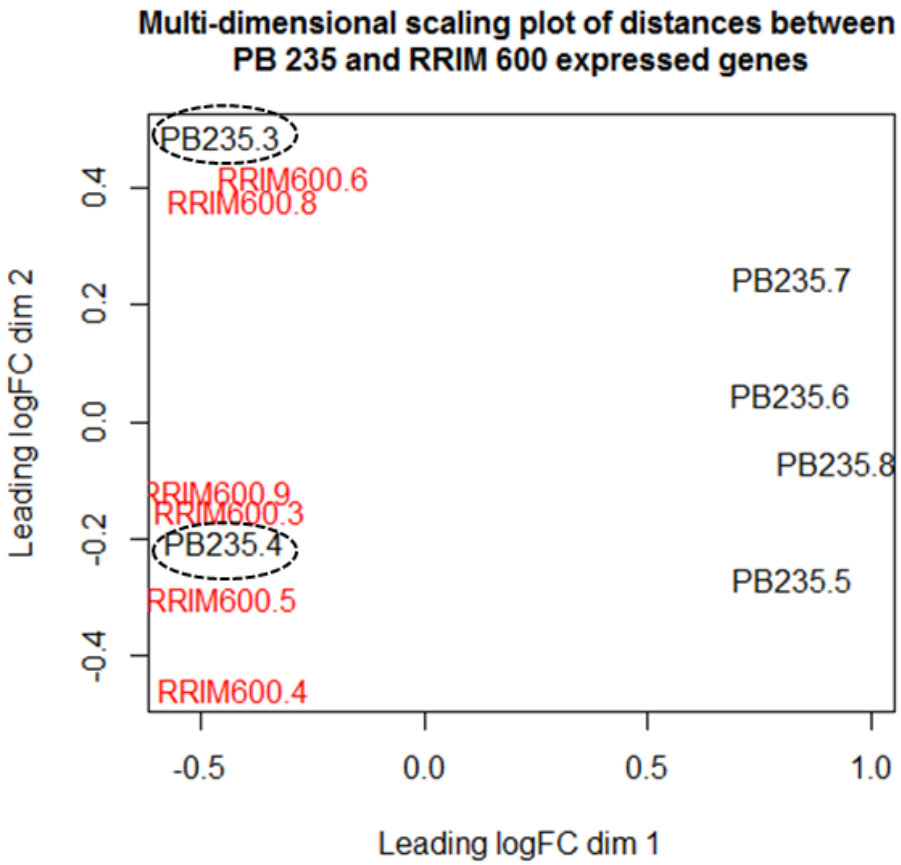


Figure 5.2.4.1: Multi-dimensional scaling (MDS) plot of expressed genes from the latex of PB235 and RRIM600. The plot indicates two PB235 samples (marked in dotted circles) were outliers as they clustered within RRIM600 libraries.

To investigate a possible reason behind the variation observed in PB235 RNA-seq datasets, a set of SNPs in the RNA-seq data generated from the PB235 and RRIM600 latex samples were examined. This was to explore whether the anomalous RNA-seq data could be the result of an error in the identification of sampled trees. SNP calling was performed on individual RNA-seq datasets that were mapped to draft genome sequence (Tang et al, 2016) using Freebayes software (Garrison and Marth, 2012) as described in section 2.34, Chapter 2. A total of 267 putative SNP markers were identified from scaffold1222 and the markers clearly classified each RNA-seq sample into two

main groups, with one group contained all PB235 RNA-seq samples and the other consisted of RRIM600 samples (Figure 5.2.4.2). Interestingly, the two outlier samples (PB235.3 and PB235.4) were found to have SNP patterns that were highly similar to that of the RRIM600 genotypes. It is clearly demonstrated that the problematic PB235 samples did not exhibit SNP content expected from PB235 *Hevea* genotypes. Though they appeared to have a genotype consistent with RRIM600 this could not be fully confirmed as the genotype of other clones in the same trial was not available. Hence, these two samples were discarded from the downstream expression measurement analysis.

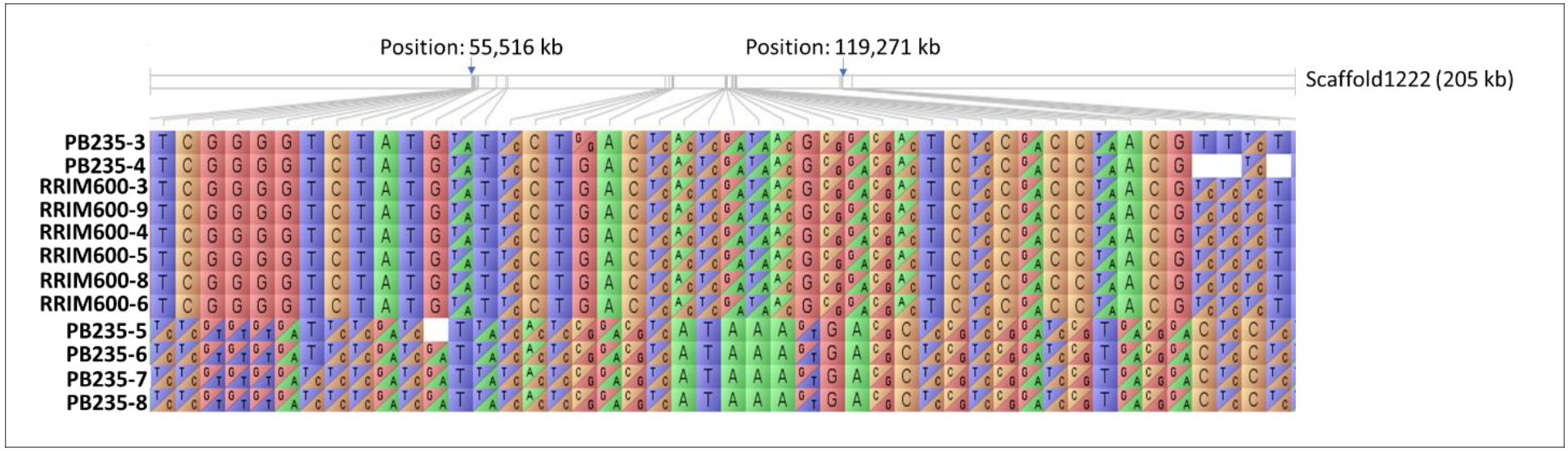


Figure 5.2.4.2: Graphical view of SNPs identified from RNA-seq data of PB235 and RRIM600 samples. The SNPs were generated based in the mapping of the cleaned reads to scaffold1222 (region 55,516 – 119,271 kb). The SNPs view was generated using Flapjack software (Milne et al., 2010).

To obtain differential expression profile of the targeted genes between PB235 and RRIM600, the mapped reads from the associated RNA-seq samples (four PB235 RNA-seq samples and six RRIM600 RNA-seq samples) were used to estimate the normalised gene expression levels. To reduce transcription noise, the genes that showed zero or $\log_2$ TPM value < 0.1 were not included in the analysis. Only genes that showed fold change of at least 1.5 ($\log_2$ TPM) and a maximum false detection rate (FDR) less than 0.001 were defined as differentially expressed. The list of differentially expressed genes is summarised in Table 5.2.4.2. From the 151 of the targeted genes, 31 were differentially expressed between PB235 and RRIM600 genotypes. These data also gave information regarding gene expression levels of individual transcripts in PB235 and RRIM600. This was illustrated by the expression plots in Figure 5.2.4.3 (a), (b) and (c).

The most notable expression patterns for the MVA, MEP pathways, isoprenoid initiator formation, rubber elongation steps and carotenoid biosynthetic steps are as follows:

i. MVA pathway: Transcript abundance of *AACT2* isoform 2 and *HMGR1* isoform 1 were higher in RRIM600 than that of PB235 (Figure 5.2.4.3(a)).
ii. MEP pathway: Gene encoding *DXS1* isoform 2 was the most abundant in both PB235 and RRIM600 latex samples. Interestingly, the gene showed higher expression level in RRIM600 in contrast to PB235 (Figure 5.2.4.3(a)).
iii. Isoprenoid initiator formation: A gene encoding *GPS* isoform 3 was found to be highly expressed in both PB235 and RRIM600. The gene level was observed to be higher in PB235 than that of RRIM600 latex. However,

the expression discrepancy between PB235 and RRIM600 was <1.5 folds, hence not classified as differentially expressed gene (Figure 5.2.4.3(b)).

iv. Rubber elongation steps: Transcript levels for *REF1, REF3, REF7. SRPP1, CPT* isoform 2 and *CPT* isoform 4 were found to be significantly expressed in PB235 and RRIM600. *REF1* showed the greatest fold change (between PB235 and RRIM600 (Figure 5.2.4.3(b)).

v. Carotenoid biosynthesis pathway: A gene encoding *PSY* isoform 2 showed the highest expression in both PB235 and RRIM600. In addition, the gene showed higher expression in PB235 (>3-fold change) than that of RRIM600 (Figure 5.2.4.3(b)).

The above genes were found to be highly expressed in latex. To investigate the extent of tissue-specific expression of these genes, analysis across multiple tissue samples from the Reyan 7-33-97 *Hevea* genotype RNA-seq datasets was also carried out. The list of tissues used for the analysis is detailed in Table 5.2.4.1. These datasets were not replicated and hence some caution should be used in interpreting the levels of differential expression. These gene expression patterns are illustrated in Figures 5.2.4.4 (a-e).

Table 5.2.4.2: List of differentially expressed genes at a fold change ≥1.5 and a maximum false detection rate (FDR) of 0.001 in the latex of PB 235 and RRIM 600 rubber tree genotypes.

	Gene identity	Raw transcript count per kilobase per million (TPM)		Expression pattern (PB 235 versus RRIM 600)
		PB235	RRIM600	
MVA pathway	AACT2_iso2	8,391	21,936	Down-regulated
	HMGR1_iso3	9	76	Down-regulated
	MK_iso2	2,492	20	Up-regulated
	MK_iso3	80	1,310	Down-regulated
	MVD1	7,141	1,509	Up-regulated
	MVD2	524	58	Up-regulated
MEP pathway	DXS1_iso2	16,345	39,975	Up-regulated
	DXS2_iso2	768	1,786	Down-regulated
Carotenoid formation	PSY_iso2	6,202	1,919	Up-regulated
	CRTISO2_iso1	72	27	Up-regulated
	LCYB_iso1	1,015	656	Up-regulated
	VDE	19	8	Up-regulated
	NCED3	147	35	Up-regulated
	NCED6	2,997	812	Up-regulated
	NCED9	18	4	Up-regulated
Isoprenoid initiator formation	IPPI_iso2	635	408	Up-regulated
	GPS_iso4	2,953	1,774	Up-regulated
	GPS_iso2	2,978	1,904	Up-regulated
	FPPS_iso1	1,039	1,936	Down-regulated
	FPPS_iso2	141	419	Down-regulated
Rubber elongation steps	CPT_iso1	714	2,975	Down-regulated
	CPT_iso3	4	30	Down-regulated
	CPT_iso5	2	13	Down-regulated
	RBIP_iso1	46,135	30,123	Up-regulated
	RBSP_iso4	6,289	18,837	Down-regulated
	REF1	160,748	283,251	Down-regulated
	REF2	1,841	542	Up-regulated
	REF3	313,082	217,051	Up-regulated
	REF4	7,547	3,397	Up-regulated
	REF5	6,752	3,827	Up-regulated
	REF8	25,848	8,786	Up-regulated

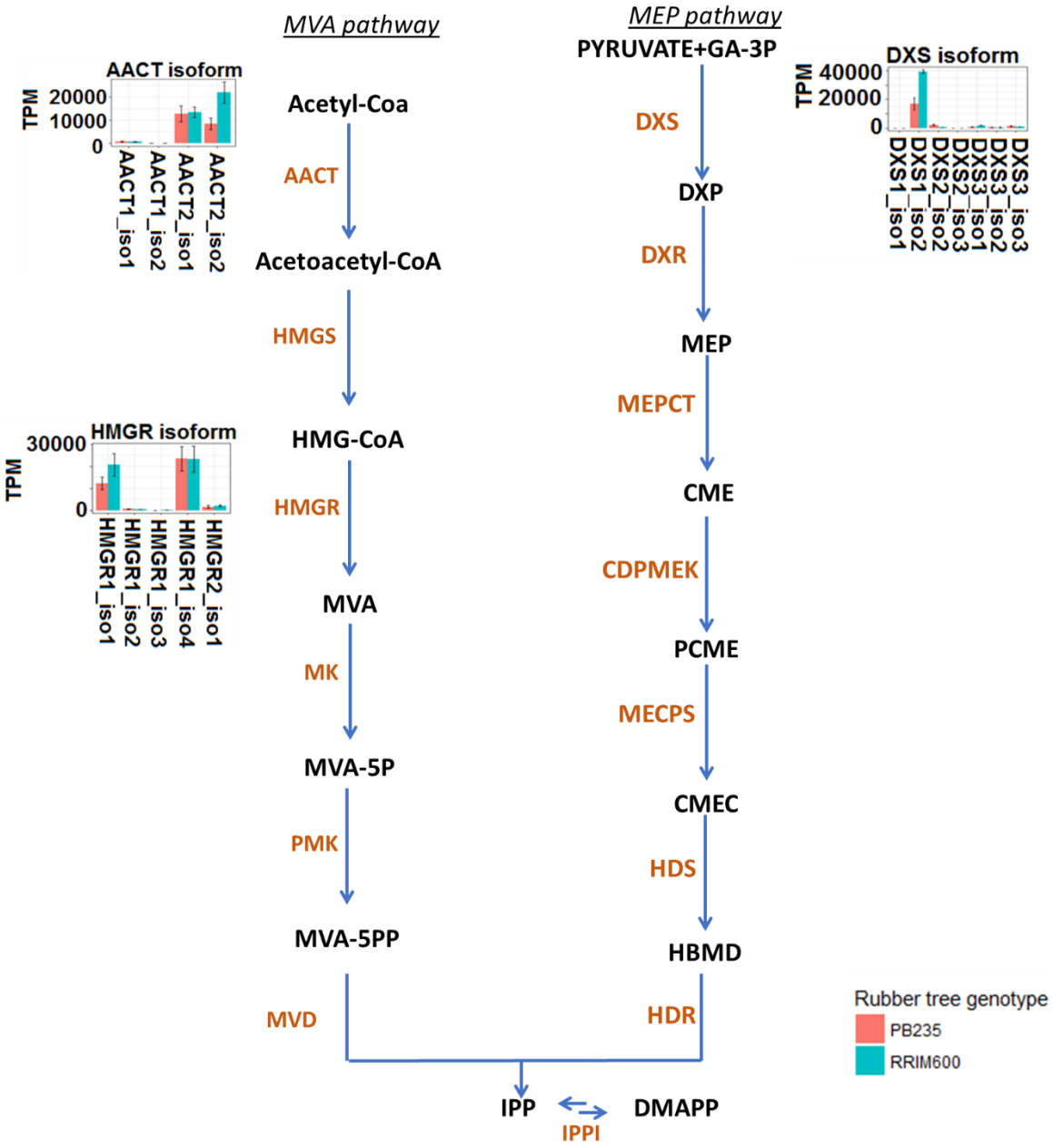


Figure 5.2.4.3(a): Schematic representation of the MVA and the MEP biosynthetic pathways. The expression level of the key genes for each step is shown the corresponding expression plot. The plot was generated based on the average of expression from RRIM600 (six replicates) and PB235 (four replicates) and the whiskers of each plot indicate standard error for each expression value.

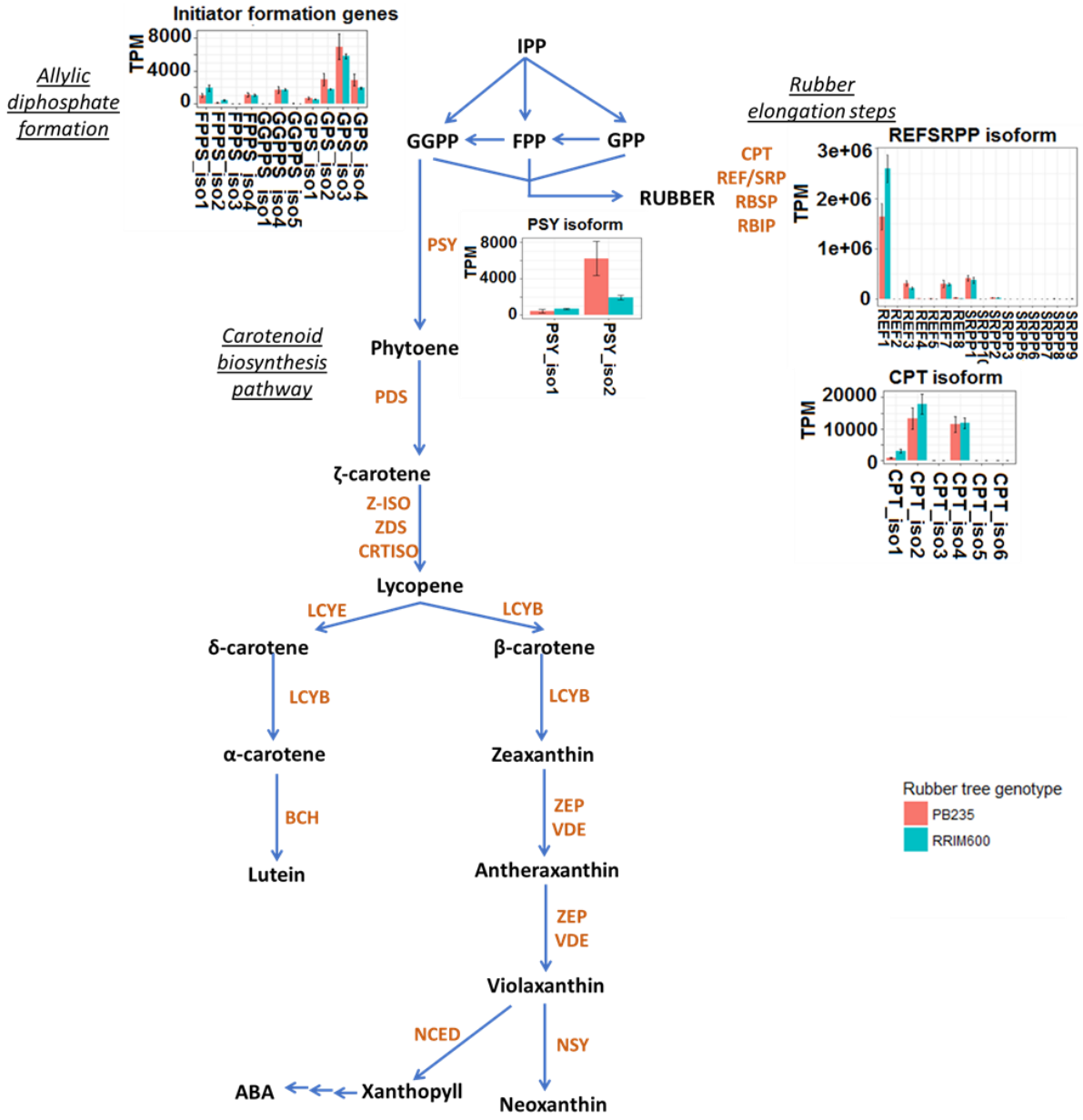


Figure 5.2.4.3(b): The steps for the formation of isoprenoid initiators (FPP, GPP and GGPP) that are used in the rubber and carotenoid biosynthetic pathways. The expression level of key genes for each step is shown . The plot was generated based on the average of expression from RRIM600 (six replicates) and PB235 (four replicates) and the whiskers of each plot indicate standard error for each expression value.

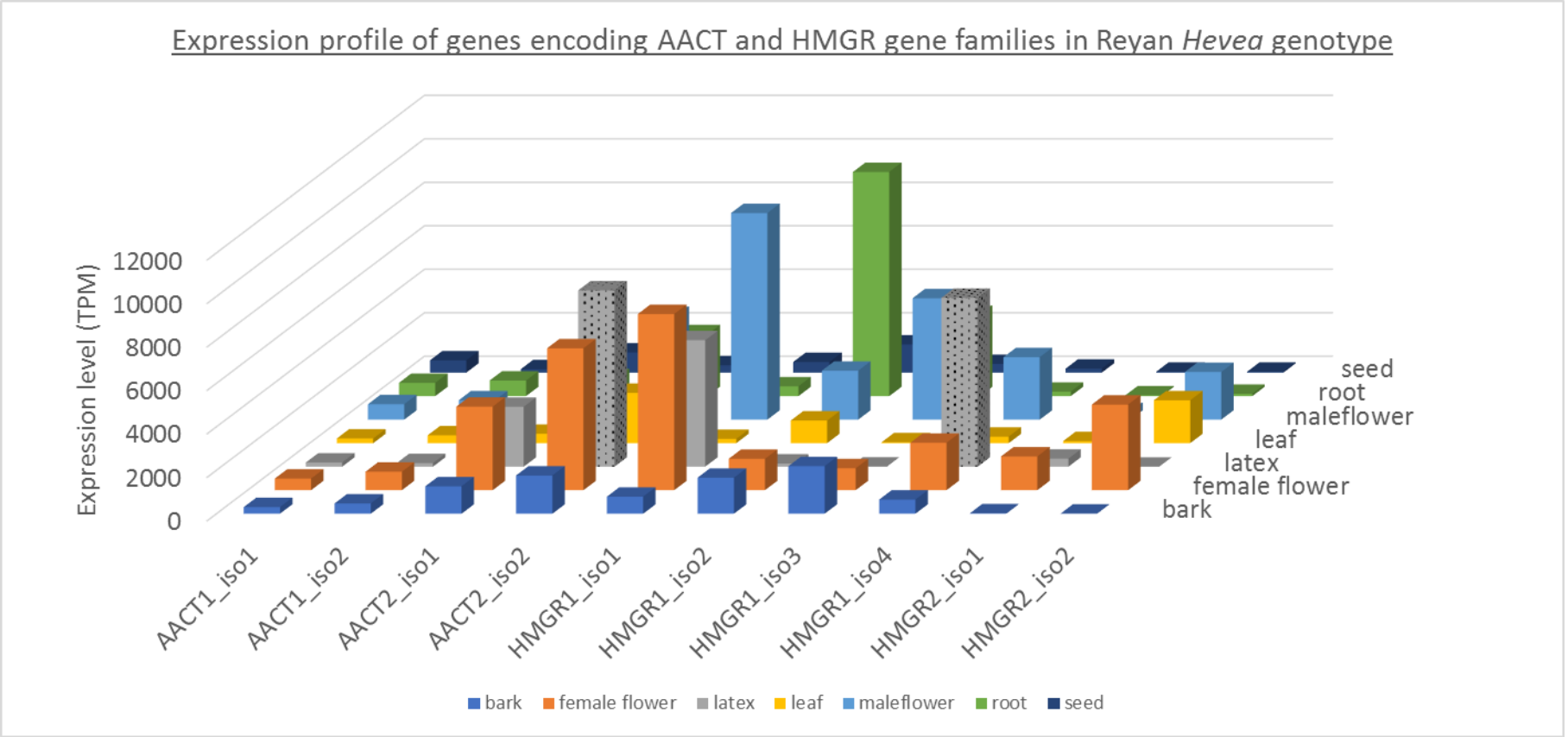


Figure 5.2.4.4 (a): Expression level of genes encoding AACT and HMGR from the MVA biosynthetic steps. The gene panel was measured from bark, latex, lead, female flower, male flower, root and seed tissues of Reyan 7-33-97 *Hevea* genotype. The dotted bars represent genes encoding AACT and HMGR isoforms that are predominantly expressed in the latex.

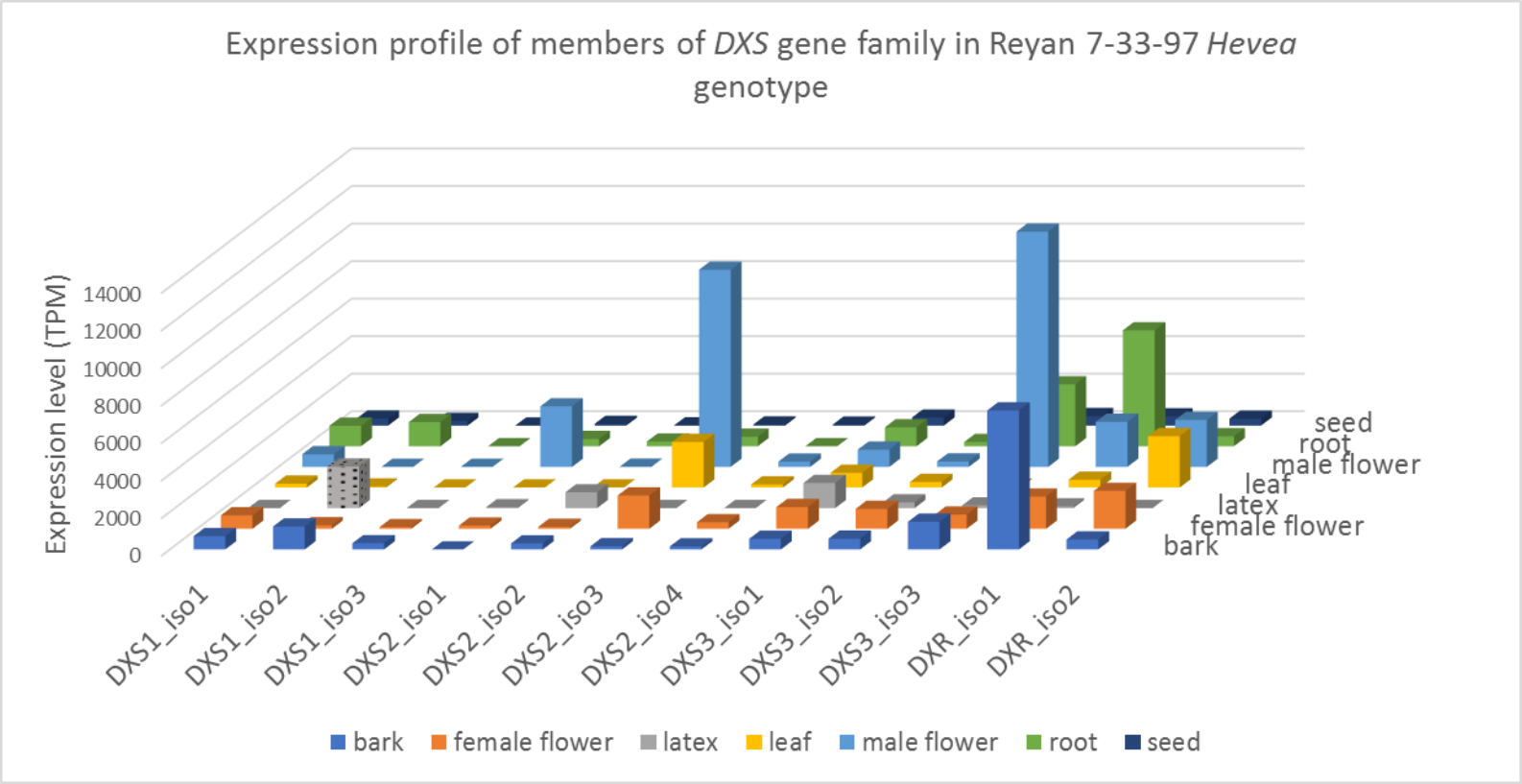


Figure 5.2.4.4 (b): Expression level of DXS gene family from the MEP biosynthesis pathway. The gene panel was measured from bark, latex, lead, female flower, male flower, root and seed tissues of Reyan 7-33-97 *Hevea* genotype. The dotted bar represents the gene encoding *DXS1* isoform 2 that predominantly expressed in the latex.

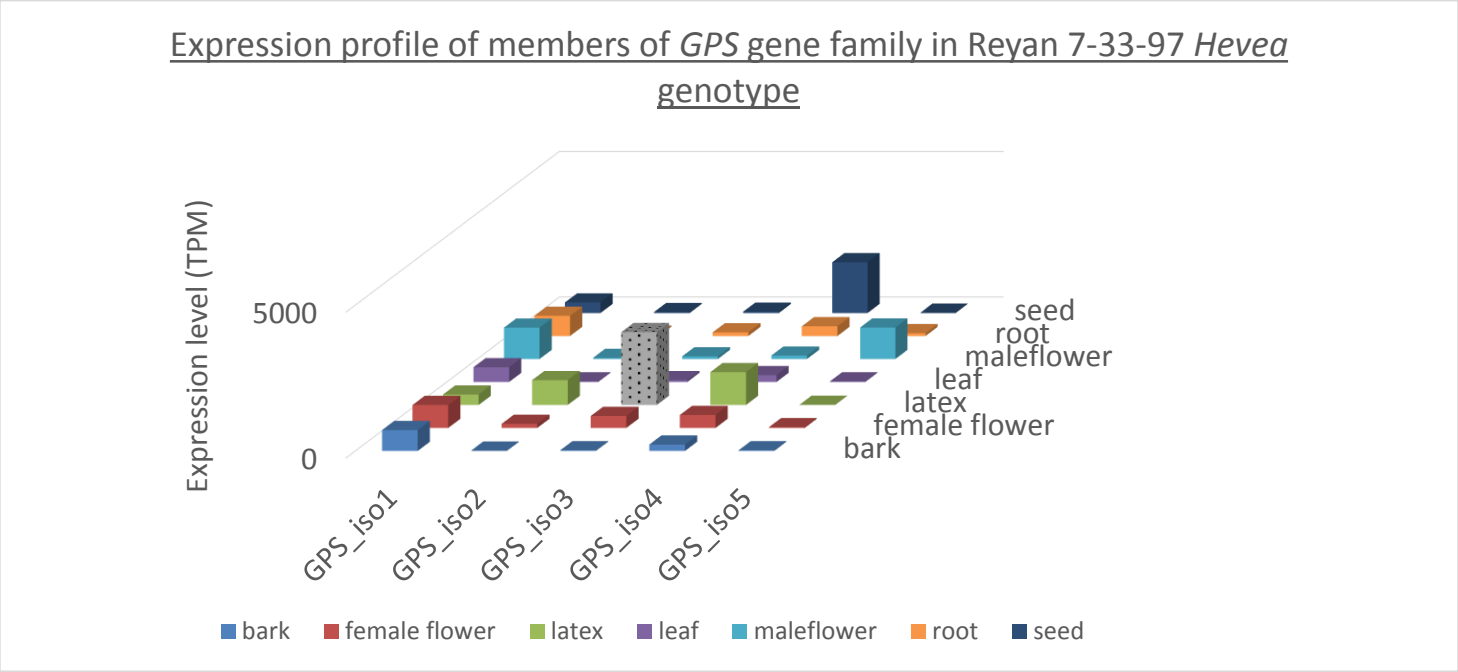


Figure 5.2.4.4 (c): Expression level of GPS gene family members, from the isoprenoid initiator formation steps. The gene panel was measured from bark, latex, lead, female flower, male flower, root and seed tissues of Reyan 7-33-97 *Hevea* tree genotype. The latex-specific isoform (*GPS* isoform 3) is labelled as a dotted bar.

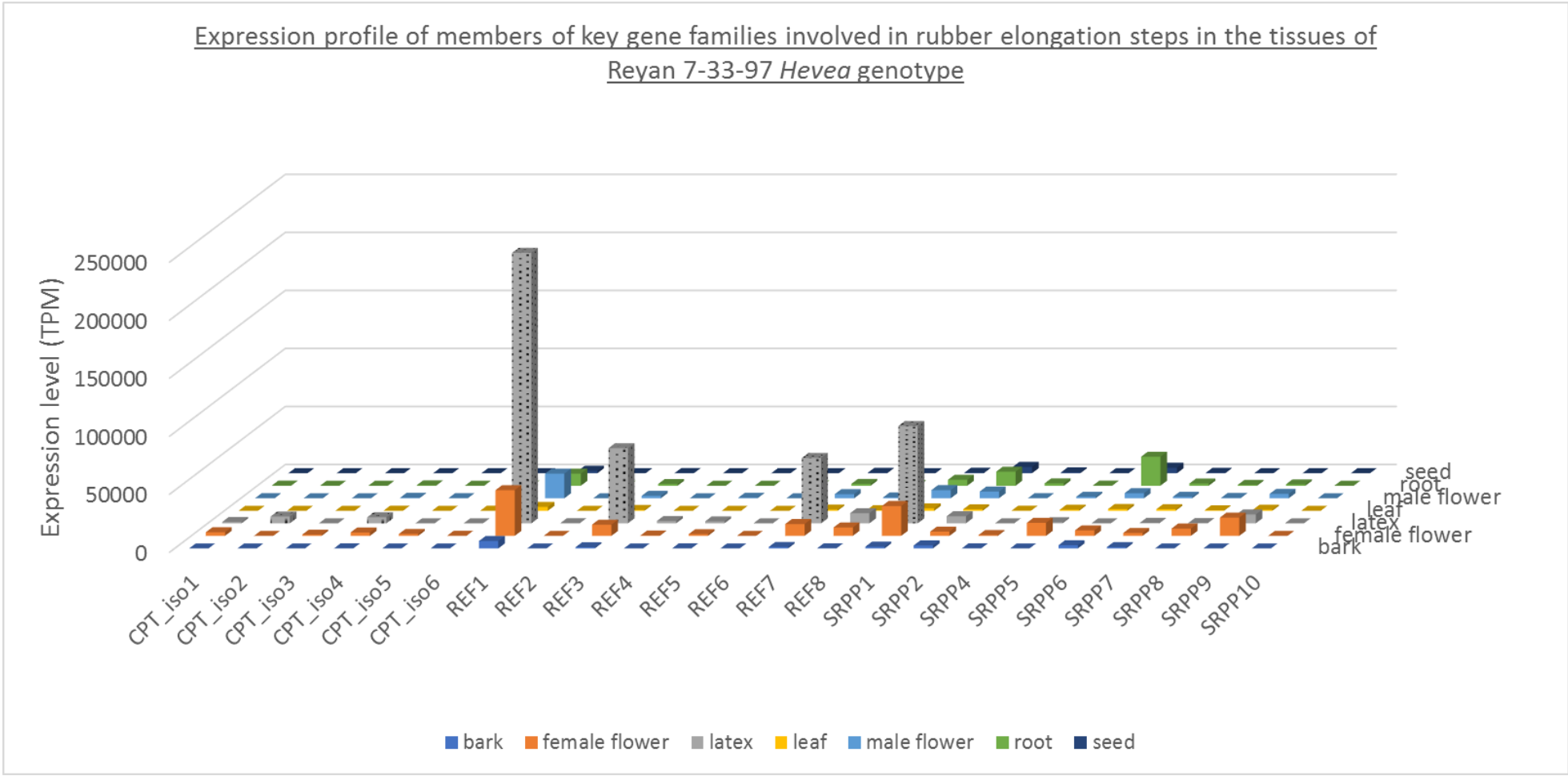


Figure 5.2.4.4 (d): Expression level of genes encoding CPT, REF and SRPP from the rubber elongation steps. The gene panel was measured from bark, latex, lead, female flower, male flower, root and seed tissues of Reyan 7-33-97 *Hevea* tree genotype. *CPT* isoform 2, *CPT* isoform 4, *REF1, REF3, REF7* and *SRPP1* that are highly expressed in latex are labelled as dotted bars.

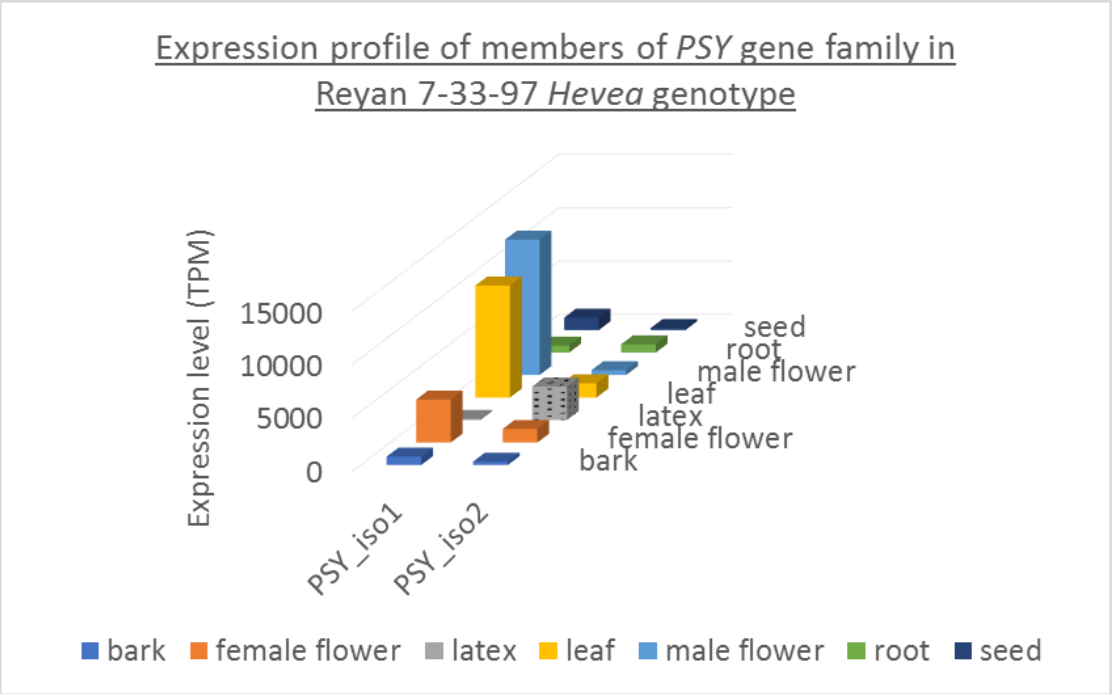


Figure 5.2.4.4 (e): Expression level of genes encoding PSY isoforms from the carotenoid biosynthetic pathway. The gene panel was measured from bark, latex, lead, female flower, male flower, root and seed tissues of Reyan 7-33-97 *Hevea* genotype. The dotted-bar representing *PSY* isoform 2 showed the highest expression in latex.

5.3. Discussion

Although four versions of *Hevea* genome assemblies have been reported (Tang et al, 2016, Lau et al, 2016, Pootakham et al, 2017, Mollison et al, 2014), these data are still at draft level, with a large number of unconnected scaffold sequences. Scaffolding inserts have interrupted a significant portion of the predicted gene model in these genome assemblies. This means that the available genome sequence is not an accurate reference on which to base the transcript repertoire. Based on the recent work in *Arabidopsis*, it is clear that the accuracy of transcript profiling is highly dependent on the quality of the reference transcript sequences (Zhang et al., 2015, Zhang et al., 2016, Brown et al., 2017b). Initially, at the conception of the current project there was an approved proposal to generate an improved reference genome based on pseudomolecule quality information, utilising long-read technology. Unfortunately, this did not come to fruition and until a new, pseudomolecular version of the *Hevea* genome is available, a reference transcriptome provides the best underlying resource for analysis of gene expression.

The utilisation of the SNP data to resolve the mis-assigned samples in the RNA samples was particularly valuable. Sampling error of this sort is a common occurrence though often regarded as unrecognised source of problems with experimental data, particularly when samples are obtained from field scale biological resources. The mis-assignment might occur in number of possible ways, such as i) at the field scale, due to errors in clonal-propagation or during planting of the trial; or ii) in the lab where it is possible to mix samples. In the present study, it is unlikely for the error to have happened in the laboratory. If the error arisen from the mishandling in the lab, we would have

found the mis-identification error in samples of both genotypes. Mis-labelling of planting materials or tracking issues has been reported in many plant species, such as acacia (Asif et al., 2017, Ratnam et al., 2017), *Arabidopsis* (Anastasio et al., 2011) and palm oil (Akpo et al., 2014). In this study, two of the PB235 samples were found to be mis-identified. The genotyping analysis based on SNP markers of the individual RNA-seq samples has confirmed that the two outlier samples were not of the PB235 rubber tree genotypes. This observation in the current project may give an indication that quality control (in the form of genotyping) of *Hevea* trials may be a valuable approach, as Yabe et al. (2018) has demonstrated that mislabeling affected the accuracy of selection prediction and overall genetic gain in breeding program.

In dealing with field planting materials, it is not uncommon to have high heterogeneity among the biological replicates. Under the field condition, the manifested plant phenotypes are controlled by the genetic content and also by environmental factors. In this study, heterogeneity of the samples can be detected from the expression pattern of the individual samples, as shown in the earlier MDS plot. Sample heterogeneity is not rare in a wide range of biological material as it has been reported and evaluated in gene network analysis from *Arabidopsis* (Uygun et al., 2016), in *Eucalyptus* germplasm maintenance (Keil and Griffin, 1994) and in the transcriptome profiling of human pancreatic β-cell (Nica et al., 2013). Despite the heterogeneity of the results, the pipeline used in the expression study was robust enough to identify a list of differentially expressed genes that were involved in the key pathways leading to the formation of rubber and carotenoid.

Although *Hevea* is an important crop for natural rubber production, there are still knowledge gaps in the uptake of molecular or genomic-assisted technique in *Hevea* breeding for enhancing the latex formation. In *Hevea* latex, the precursor for rubber formation, IPP is generated from both the plastidic MEP pathway and the cytosolic MVA pathway. Apart from rubber formation, IPP is also utilised in the formation of carotenoid in *Hevea* latex. Therefore, gaining insights into the regulation of key genes involved in rubber and carotenoid formation in general could accelerate breeding activity for an enhanced latex production. In this study, all gene families in the pathways were identified from the reference transcriptome. For each gene family, higher resolution of alternatively spliced transcripts was attained and such transcripts may represent different parts of a single gene, different members of a gene family, or both.

To gain information on the synthesis of rubber and carotenoids in the latex of *Hevea brasiliensis*, differential gene expression based on the transcriptomic data was analysed. In the previous Chapter 3, dry rubber content and carotenoid levels were also measured from the same latex samples. From the metabolite measurement analysis, the rubber content was found to be higher in RRIM600 while the accumulation of carotenoid was greater in PB235. Ideally, a correlation between the gene expression and the metabolite accumulations should be performed for a deeper understanding of the IPP utilisation between rubber and carotenoid formation. However, the limited number of replicates in this study will not allow for a meaningful correlation of the transcript levels to that of rubber content and carotenoid accumulation in *Hevea* latex. Nevertheless, patterns of the transcript levels, rubber content and carotenoid quantity can be discerned in this study. For example, it has been shown that genes encoding AACT, HMGR from the MVA pathway were found

be expressed at a higher in RRIM600 compared to that of PB235. Indeed, high HMGR activity in *Hevea* latex was first detected by Hepper and Audley (1969), and the enzyme was reported to be the rate-limiting factor in the isoprenoid biosynthesis pathway (Chappell, 1995). Later, the transcript for *HMGR1* was detected to be preferentially expressed in the laticifer (Sando et al., 2008a, Venkatachalam et al., 2007, Ji et al., 1993). In the latest study, the elevated expression of *HMGR1* gene in 6-year old transgenic *Hevea* was reported to show an increased amount of rubber content when compared to that of the non-transgenic *Hevea* (Jayashree et al., 2018). While many findings have indicated the importance of *HMGR1* in the MVA pathway, AACT gene has not been always been given the same importance. However, in other non-*Hevea* plant that also producing latex known as *Taraxacum koksaghyz*, *AACT2* was inferred to play a role in rubber formation as its overexpression was reported to be positively correlated to the increased latex yield (Pütter et al., 2017).

Interestingly, one of DXS genes, *DXS1* isoform 2 was found to be significantly high (> 2-fold change) in the RRIM600 latex in contrast to the PB235 latex. Chow et al (2012) also reported the same expression pattern in a separate study and they postulated that the MEP pathway may act as an alternative IPP provider for rubber formation, rather than involved in the synthesis of carotenoid in the latex of mature *Hevea* RRIM600. The lower accumulation of carotenoid in the latex of RRIM600 and higher expression of *DXS1* isoform 2 in this study has provided an additional support to the suggestion that the overall activity of the MEP pathway in RRIM600 may provide the IPP precursor to the rubber biosynthesis.

Cornish and Blakeslee (2011) has reported that the activity of rubber transferase in rubber formation is positively correlated to the pool of isoprenoid initiator (FPP, GGPP and GPP). Therefore, it is expected that gene families involved in the formation of the isoprenoid initiators might be high in the latex of *Hevea* brasiliensis. Surprisingly, it was observed that transcripts for the key gene families involved in the formation of isoprenoid initiators (FPPS, GPPS and GPS) were generally low. This suggest that the genes may be subjected to a post-translation modification as its low expression levels do not correlate with the high content of rubber in the RRIM600 samples.

The genes for REFSRPP and CPT were observed to be the most highly expressed in both PB235 and RRIM600. This is not surprising as Yamashita et al. (2016), Brown et al. (2017a) have showed that interaction of these proteins is required for rubber formation *in vitro.* Of the members of the REFSRPP gene family, *REF1* showed the most significant differential expression between RRIM600 and PB235 (>1.5-fold change). The preferential expression of *REF1* in *Hevea* trees with high latex production has also been reported in the past. Venkatachalam et al. (2007) demonstrated that *REF1* expression was at an average of 5-fold higher in high latex yielding genotypes, compared to that of the low latex yielding genotypes. In addition to *REF1,* other gene members namely *SRPP1*, *REF3* and *REF3* also showed high expression in latex tissue.

There were eight isoforms detected for REF and ten for SRPP identified from the reference transcript database. On the other hand, from the 11 CPT genes annotated in the *Hevea* draft genome, only 6 CPT were defined from the reference transcript database. These genes also exhibited different regulation for each isoform, whereby REF1, REF3, SRPP1, CPT isoform 2 were highly

expressed in the latex samples collected from PB 235 and RRIM 600 genotypes. However, REF1 and CPT isoform 2 are significantly higher in RRIM 600 compared to that of PB 235. This is the first data that demonstrated the transcript-specificity of these genes and co-expression of these gene members in the latex of *Hevea brasiliensis*. A possible explanation to account for the differential regulation of REF, SRPP and CPT gene families is that their genes have been expanded after through evolution of the *Hevea* genome (Tang et al, 2016, Lau et al, 2017). The duplication event has caused these genes to undergo sub-functionalisation in latex tissue. For example, not all REF and SRPP isoform are proposed to be involved in elevating the rubber biosynthesis activity. Instead, only REF1 and SRPP1 were reported to display such function (Oh et al, 1999; Dennis and Light, 1989). Indeed, other REF and SRPP isoforms (REF8 and SRPP10) were implicated in the maintenance of rubber particle structure (Dai et al., 2017). For CPT, there was a report suggested that CPT isoform 2 involved in the synthesis of rubber chain (Asawatreratanakul et al., 2003). Currently, there are no studies on the expression or function for other CPT gene members in *Hevea*. However, its orthologs from *Arabidopsis* were demonstrated to involve the formation of short-chain of polyisoprenoid alcohols (Akhtar et al., 2017).

The key genes identified in other species related to the carotenoid biosynthesis pathway (Hirschberg, 2001) were found to be up-regulated in PB235. This is expected as carotenoid accumulation in PB235 was found to be higher than that of RRIM600. In particular the PSY enzyme which catalyses the first committed step in the carotenoid biosynthetic pathway (Sun et al., 2018) and has been found to be the rate-limiting factor in the accumulation of carotenoid in other species (Sun et al., 2018, Alagoz et al., 2018). While PSY is encoded by a multi-gene family, not all PSY gene members play a role in

carotenoid biosynthesis. PSY gene member universally known as *PSY1* in many plant species has been implicated in carotenoid accumulation in the flesh tissue of tomato fruit (Kachanovsky et al., 2012); *PSY2* (identified as *PSY* isoform 2 in this study) was implicated in the modulation of carotenoid formation in the vegetative tissue of apple (Ampomah-Dwamena et al., 2015) and banana (Dhandapani et al., 2017); while *PSY3* was reported to play a role in rice endosperm (Welsch et al., 2008). From *Hevea*, transcripts (identified as PSY isoform 1, PSY isoform 2, PSY isoform 3 and PSY isoform 4) with high similarity (>90%) to *PSY2* were detected. None of transcripts related to *PSY1* and *PSY3* were identified from the genomic and transcriptome resources. With regard to its role in modulating carotenoid accumulation, it was demonstrated that increased levels of PSY gene expression have enhanced the formation of β-carotene in multiple plant systems such as *Brassica*, potato, cassava and tomato (Shewmaker et al., 1999, Giuliano et al., 2008, Welsch et al., 2010, Morris et al., 2004, Ducreux et al., 2005). These results are consistent with the present study in which β-carotene was found to be the major *Hevea* carotenoid, responsible for the undesirable yellow colour in latex. In addition, increased levels of PSY gene expression were found in the high carotenoid genotype. Further work may be required to determine whether this is a direct effect of the PSY gene or the result of regulation by other transcriptional regulatory components (i.e. transcription factor or promoter). If a direct result of variation in the PSY gene levels can be confirmed, it may be possible to identify SNP markers associated with low latex carotenoid levels which can be deployed in breeding activity.

As a conclusion, this is the first attempt to develop a reference transcriptome in *Hevea* constructed from multiple sequencing platform. The

transcriptome analysis has shown latex-specificity expression pattern of key genes implicated in the formation of rubber and carotenoids. In addition, the expression patterns of latex specific isoforms are related to latex and beta carotene accumulation. These findings highlighted the utility of the reference transcriptome in profiling expression changes within *Hevea* latex.

Chapter 6

Characterisation of REFSRPP gene family

6.1.1. Brief Introduction

In the previous chapter, the utility of the *Hevea* reference transcriptome was demonstrated through the differential expression analysis of key genes involved in rubber and carotenoid formation. One of the key findings from the analysis indicated that genes encoding rubber elongation factor (REF) and small rubber particle protein (SRPP) were highly expressed in the latex tissue of *Hevea brasiliensis*. In this chapter, subsequent work in characterising the REF and SRPP genes was undertaken, by integrating the information generated from the merged transcriptome and *Hevea* draft genome.

6.1.2. REF and SRPP gene family

From the histochemical and proteomic studies, REF and SRPP were found to be predominantly localised on the rubber particle surface (Singh et al., 2003, Bahri and Hamzah, 1996, Wang et al., 2018). Initially, when a predominant rubber particle protein was first detected and found to play a role in rubber chain elongation, Dennis et al. (1989) has called the protein as REF. Later, when another prominent protein associated with small rubber particle was characterised, Oh et al. (1999) has termed the protein as SRPP. These two proteins share about 70% sequence similarity and was found to carry a domain known as REF domain (Oh et al., 1999, Sookmark et al., 2002). When REF was observed to be predominantly present on the surface of large rubber particles and SRPP seemed to preferentially occur on the small rubber particle protein, the nomenclature of the proteins seemed to be on the basis of the protein localisation on the rubber particles (Yeang et al., 1996). However, it has since

confirmed that not only is REF ubiquitously available on the large rubber particles, the protein also found to be associated with the small rubber particles (Bahri and Hamzah, 1996, Berthelot et al., 2014a, Berthelot et al., 2014b).

After more REF and SRPP gene members were identified by Chow et al. (2007), they suggested that both REF and SRPP are of two sub-proteins and collectively form a large gene family known as rubber particle membrane proteins. Subsequently, eight REF genes (REF1 – REF8) and ten SRPP genes (SRPP1 – SRPP10) were identified from the *Hevea* draft genome (Tang et al., 2016). With these additional gene members, a systematic nomenclature system is needed in order to define SRPP and REF. This is because while both proteins showed a certain degree of protein similarity, they also demonstrated a different mode of interactions with the latex lipid contents and rubber particles (Berthelot et al., 2012, Berthelot et al., 2014b, Wadeesirisak et al., 2017). Yamashita et al (2016) and Brown et al (2017) have also shown REF and SRPP have different interaction with the rubber transferases in mediating the rubber formation. Thus, based on these findings, REF and SRPP are considered to be of different protein classes. At the moment, there is no way to define these two protein classes apart and in the present study, the genes encoding REF and SRPP were grouped into a single family, termed as REFSRPP gene family.

The REFSRPP gene family was also reported in other rubber-producing plants such as *Taraxacum*, lettuce, guayule, sunflower and *Ficus* (Cornish et al., 1999, Singh et al., 2003, Tanaka, 1985, Ponciano et al., 2012, Pütter et al., 2017). *Hevea brasiliensis* is a perennial crop and has a long immaturity period. Thus, in contrast to *Taraxacum* that provides a good system for a gene modification study (Pütter et al., 2017, Laibach et al., 2015, Laibach et al.,

2018), any gene modification analysis in *Hevea brasiliensis* will not be able to generate data within a short time frame. Therefore, most of the confirmations regarding the role of REF and SRPP genes in elevating rubber formation *in vivo* were reported in other rubber-producing plant namely dandelion and lettuce (Schmidt et al., 2010a, Collins-Silva et al., 2012, Chakrabarty et al., 2014, Laibach et al., 2018). Through functional characterisations of the REFSRPP gene family in dandelion and lettuce, it has been demonstrated that not all gene members were involved in rubber biosynthesis (Chakrabarty et al., 2014, Laibach et al., 2015). For example, for the five REFSRPP gene family members identified from *Taraxacum* (*TkSRPP1 – TkSRPP5*), only one gene member known as TkSRPP3 affected rubber accumulation and its rubber molecular weight in *Taraxacum* (Hillebrand et al., 2012). Likewise, in lettuce, eight REFSRPP gene members were identified (*LsSRPP1 – LsSRPP8*), but the transgenic lettuce with silenced individual *LsSRPP4* and *LsSRPP8* did not influence the rubber formation (Chakrabarty et al., 2014).

Meanwhile, the evidence regarding the role of REF and SRPP proteins in rubber formation in the latex of *Hevea brasiliensis* were accumulated based on rubber biosynthesis *in vitro* (Lynen, 1967, McMullen and McSweeney, 1966, Wititsuwaannakul et al., 2004, Chiang et al., 2014). Alternatively, the role of REF and SRPP in increasing the rate of rubber formation can be inferred based on their gene levels in rubber tree genotypes with differential latex yield. For instance, expression profiling for REFSRPP gene members (*REF1, REF3, REF7, REF8, SRPP1* and *SRPP2*) in three *Hevea* genotypes with differential latex production was carried out by (Tong et al., 2017). They had observed that SRPP1 and REF7 protein levels were increased in the rubber genotypes with higher latex production. This indicates that REFSRPP genes may have

undergone evolutionary changes that created multiple copies of REFSRPP genes (Tang et al., 2016, Lau et al., 2016). Therefore, detailed studies of these isoforms in relation to its genomic locations will provide an opportunity to examine the connection between REFSRPP gene and rubber production. Furthermore, the identification of polymorphism between REFSRPP isoforms will help in future breeding efforts for rubber tree.

6.1.4 Aims

The present study describes the characterisation of REFSRPP gene family based on sequence search in the *Hevea* draft genome and the previously constructed reference transcriptome. Notably, the utilisation of the merged transcriptome data in confirming the coding region of REFSRPP has improved the annotation of the REFSRPP gene models. The expansion of REFSRPP gene was observed from the phylogenetic tree construction. The REFSRPP gene expansion was further demonstrated through the comparison of REFSRPP location on *Hevea* genome scaffold to the homologous genomic region of from cassava. In addition, a pilot study on the genomic region containing REFSRPP gene clusters has indicated that these regions might have a link to the ability of *Hevea* tree producing a large amount of latex.

6.2. Results

6.2.1. REFSRPP gene family analysis

Until the publication of the Reyan 7-33-97 draft genome, it was difficult to obtain a clear picture of the size and diversity of the REFSRPP gene family in *Hevea*. This was a consequence of the extensively fragmented assembly with many genes either split between contigs or padded with scaffolding N's. However, the availability of the much higher quality Reyan 7-33-97 assembly together with the availability of extensive reference transcriptome have allowed a much more coherent picture to emerge. The more robust framework based on the Reyan 7-33-97 genome also allows a re-assessment of the data from the more fragmented assemblies. Currently, a systematic classification that could identify REF and SRPP genes is yet to be proposed. Therefore, the present study has grouped all REF and SRPP sequences under a common name of REFSRPP gene family.

Based on the TBLASTN sequence search, 18 REFSRPP genes were identified from the Reyan 7-33-97 draft genome. The gene identifiers (*REF1 – REF8* and *SRPP1 – SRPP10*) reported by Tang et al (2016) were retained for the REFSRPP genes investigated in this study. The 18 REFSRPP genes were scattered across six *Hevea* genome scaffolds, as illustrated in Figure 6.2.1.1. Of these, 12 REFSRPP genes were tandemly arranged on scaffold1222 while other scaffolds (scaffold0818, scaffold2358, scaffold0197, scaffold0824, scaffold0624 and scaffold0916) harbour a single REFSRPP gene. Based on the previous expression analysis of REFSRPP gene family (Chapter 5, section 5.2.3), the REFSRPP gene members with the highest expression level (*REF1, REF3, REF7* and *SRPP1*) was found to be located on scaffold1222.

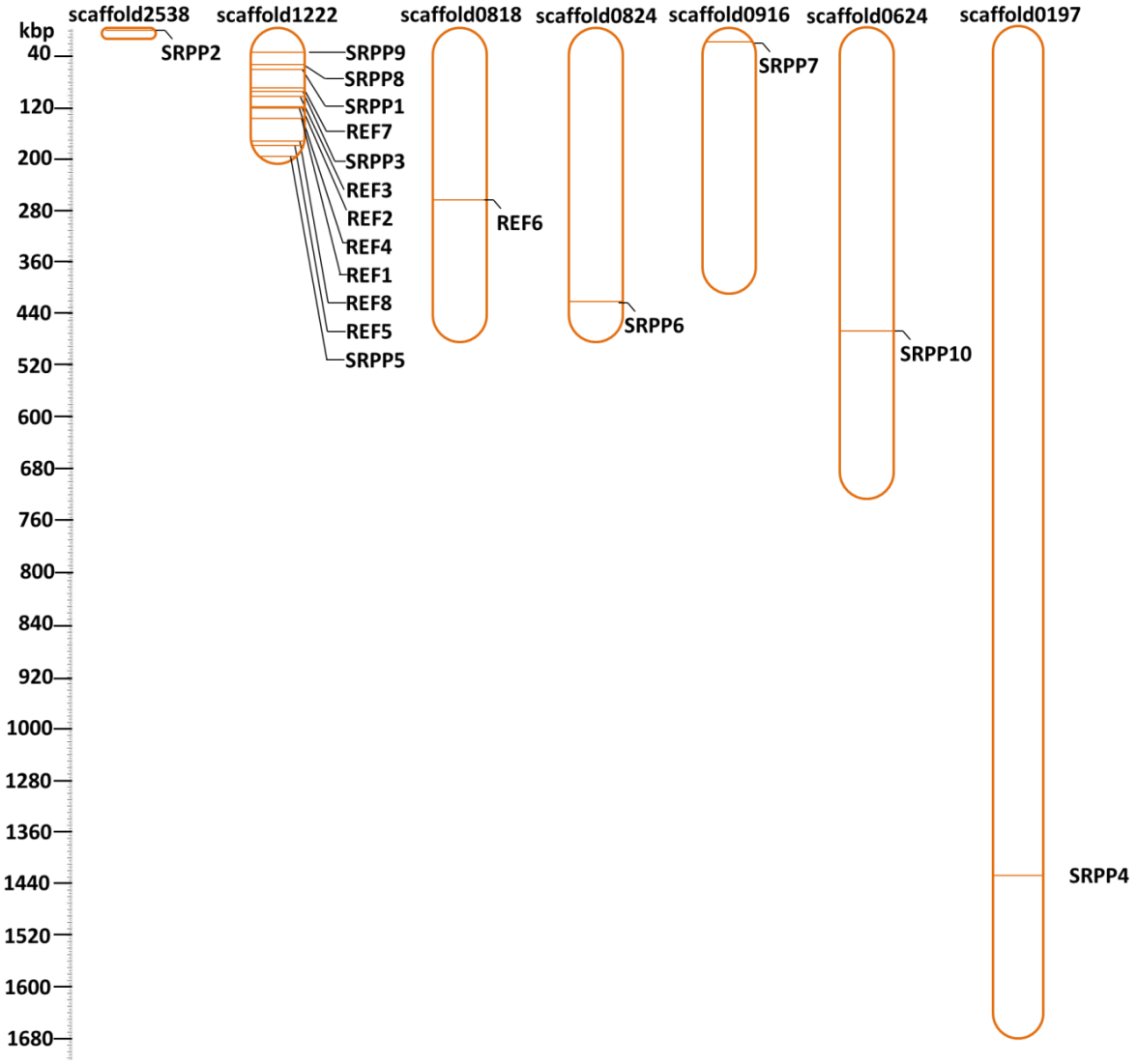


Figure 6.2.1.1: The location of REFSRPP genes annotated on the scaffolds assembled from Reyan 7-33-97 genome sequence.

Once the REFSRPP gene models were confirmed based on the support from the reference transcriptome, it also emerged that these gene models were detectable in the other *Hevea* draft genomes based on TBLASTN analysis (RRIM928, BPM24 and RRIM600) (Table 6.2.1.1). This indicated that there is no variation in the REFSRPP gene family composition among the *Hevea* genotypes. The reference transcriptome provided support for the exon-intron boundaries and information on the number of the predicted transcript variants. Introns for REFSRPP genes all displayed GT (or CG in some limited cases) at the 5’ splice site and AG at the 3’ splice site, which are the canonical splicing sites in plant introns (Brown et al., 2017b). Likewise, the upstream (5’) and downstream (3’) untranslated regions for the REFSRPP gene models were also extended based on the support from the reference transcriptome.

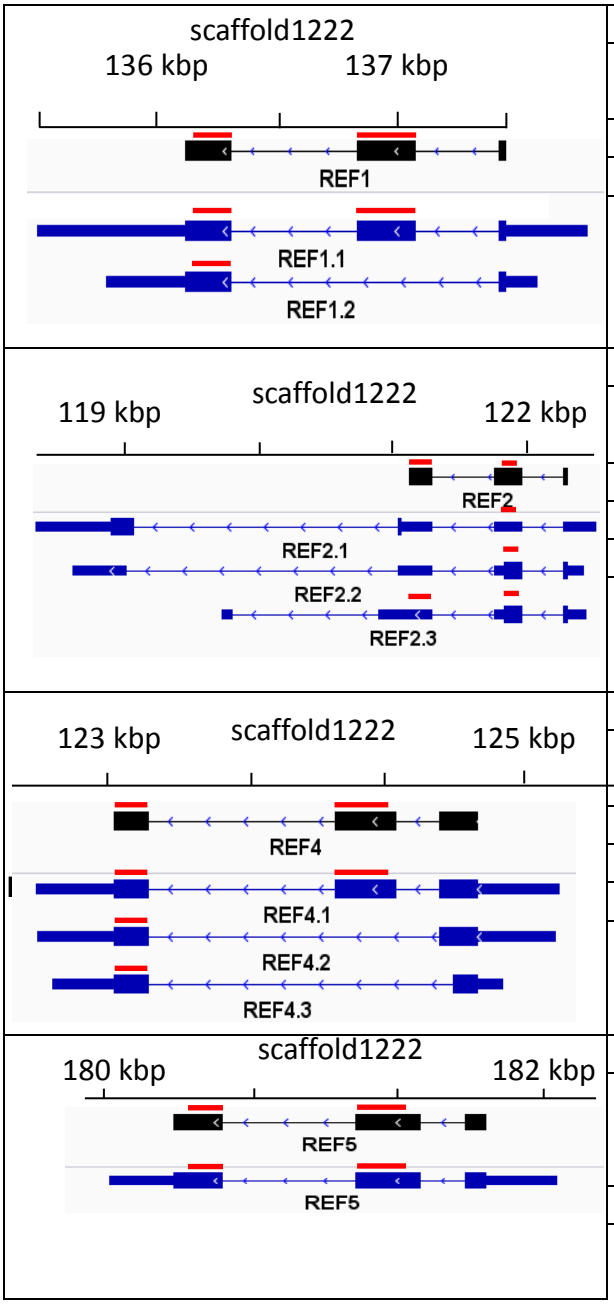
All REFSRPP sequences were observed to have a conserved region known as REF domain (Pfam Accession No. 05755) (Figure 6.2.1.2). REF domain is encoded by an average of 110 amino acid residues (Tong et al., 2017) and resides over two REFSRPP exons (Table 6.2.1.2). The sub-group between REF and SRPP sequences can be further corroborated by the disparity of their carboxyl-terminal (C-terminal) region. A common feature shown by REF sequences is a shorter C-terminal region, while said region is generally longer in SRPP sequences. This pattern is consistent with the earlier observations made on REF1, SRPP1, REF7 and SRPP8 gene members (Berthelot et al., 2014a, Tong et al., 2017). Therefore, the C-terminal distinction can be used to define the REF and SRPP protein classes. Indeed, *SRPP8, SRPP9* and *SRPP10* did not exhibit the extended N-terminal region and therefore, the re-classification of these sequences as REF instead of SRPP is

strongly recommended. However, for consistency, the sequence identifier proposed by Tang et al was still employed in the present study.

Table 6.2.1.1: REFSRPP genes identified from reyan, BPM24, RRIM 600 and RRIM 928 draft genomes.

	REF/SRPP genes identified from the *Hevea* draft genome sequences (predicted protein length)			
	Reyan	**RRIM 600**	**BPM 24**	**RRIM 928**
REF1	scaffold1222_136753 (138)	Contig149683_2792 (117)	BDHL01043746_305800 (80)	C34965234_1 (82)
REF2	scaffold1222_121702 (139)	Contig1741_29448 (129)	BDHL01043746_288500 (139)	C34965234_1 (82)
REF3	scaffold1222_100110 (175)	Contig1741_29791 (175)	BDHL01043746_277500 (130)	C33935892_1 (67)
REF4	scaffold1222_124165 (164)	Contig111648_1 (163)	BDHL01043746_296267 (130)	scaffold64928_19000 (45)
REF5	scaffold1222_181260 (154)	Contig345_51060 (154)	BDHL01040996_361900 (154)	scaffold7254_17100 (152)
REF6	scaffold0818_272553 (77)	Contig152868_20500 (77)	BDHL01030548_473280 (77)	scaffold281693_8000 (62)
REF7	scaffold1222_89338 (117)	Contig1741_73500 (117)	BDHL01043746_266000 (117)	scaffold18376_27700 (117)
REF8	scaffold1222_175215 (222)	Contig345_61500 (209)	BDHL01043746_223500 (282)	scaffold7254_9500 (189)
SRPP1	scaffold1222_60641 (204)	Contig1741_118045 (204)	BDHL01043746_229400 (204)	scaffold18376_58700 (76)
SRPP2	scaffold2538_3915 (243)	Contig617_192536 (243)	BDHL01002762_99500 (243)	scaffold48130_64000 (243)
SRPP3	scaffold1222_95474 (204)	Contig1741_66500 (152)	BDHL01043746_272400 (197)	scaffold18376_18600 (205)
SRPP4	scaffold0197_1429806(230)	Contig24714_65348 (230)	BDHL01010403_149000 (214)	scaffold91764_65800 (177)
SRPP5	scaffold1222_196376 (216)	Contig345_83770 (216)	BDHL01040996_345000 (216)	scaffold7254_34200 (216)
SRPP6	scaffold0824_400587 (230)	Contig2078_181730 (230)	BDHL01017696_766300 (230)	scaffold65897_125600 (230)
SRPP7	scaffold0916_24536 (223)	Contig635_143193 (223)	BDHL01017696_275800 (223)	scaffold97371_4900 (202)
SRPP8	scaffold1222_55336 (153)	Contig1741_123030 (469)	BDHL01043746_224500 (153)	scaffold18376_63000 (152)
SRPP9	scaffold1222_37173 (189)	Contig1741_144017 (189)	BDHL01043746_ 202000 (177)	scaffold129657_1 (55)
SRPP10	scaffold0624_516697 (148)	Contig2016_43950 (148)	BDHL01010403_150100 (148)	scaffold78971_20200 (150)

The transcriptome reference has provided evidence on the REFSRPP transcript variants (Figure 6.2.1.2). It was observed that some of these transcript variants carry truncated REF domain and their expression level was found to be correlated to the intactness of the domain. The variants with an intact REF domain were observed to have a higher level of expression compared to those with a truncated version of the domain. For example, amongst the two transcript variants of *REF1*, the one with an intact REF domain (*REF1.1*, refer to Figure 6.2.1.2.) was significantly higher compared to the expression of *REF1.2* that lacked some part of REF domain. It was observed that majority of the REFSRPP gene members have a single dominant transcript (transcript variant with the highest expression levels). Apart from the conserved REF domain, the protein feature of REFSRPP gene members was predicted by identifying transmembrane region that might infer their function. The identification of the transmembrane region for the REFSRPP gene members is shown in Figure 6.2.1.3.



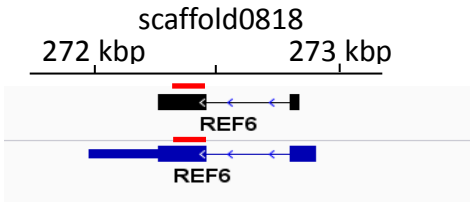
REF1		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
REF1.1	1,613,033	2,572,590
REF1.2	10,961	23,771

REF2		
Transcript variants	**Expression level (TPM)**	
	PB235	**PB235**
REF2.1	49	52
REF2.2	169	28
REF2.3	1,448	446

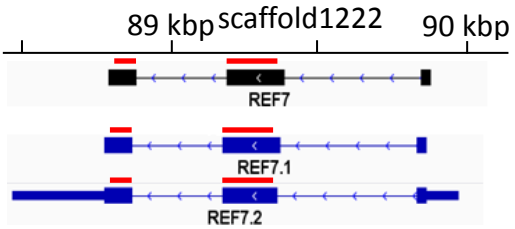
REF4		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
REF4.1	3,899	1,386
REF4.2	1,952	175
REF4.3	1,860	1,457

REF5		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
REF5	5,970	3,626

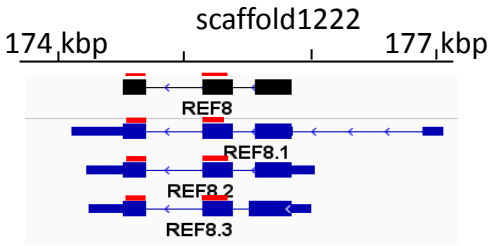
Figure 6.2.1.2: The gene models for 18 REFSRPP isoforms, with their corresponding transcript variants. The annotated gene model from genome assembly is represented by black rectangle, while the transcripts variants are represented by blue rectangle. The thicker rectangles represent the protein coding region whilst the thinner rectangles represent the untranslated coding region (UTR). The expression level for each transcript variant was calculated based on the mapped RNA-seq reads generated from the RRIM600 and PB235 latex samples in the present study. The REF domain featured in the gene models and transcript variants is denoted in red line.



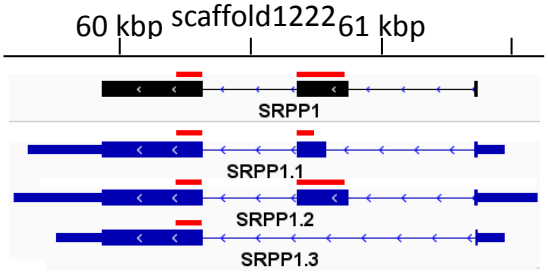
REF6: No expression data



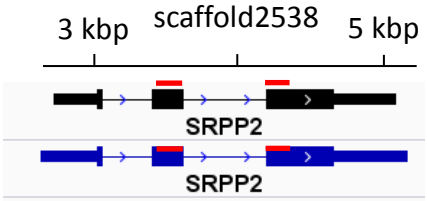
REF7		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
REF7.1	154,034	159,456
REF7.2	111,630	121,665



REF8		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
REF8.1	112	4,139
REF8.2	25,143	755
REF8.3	240	3,070

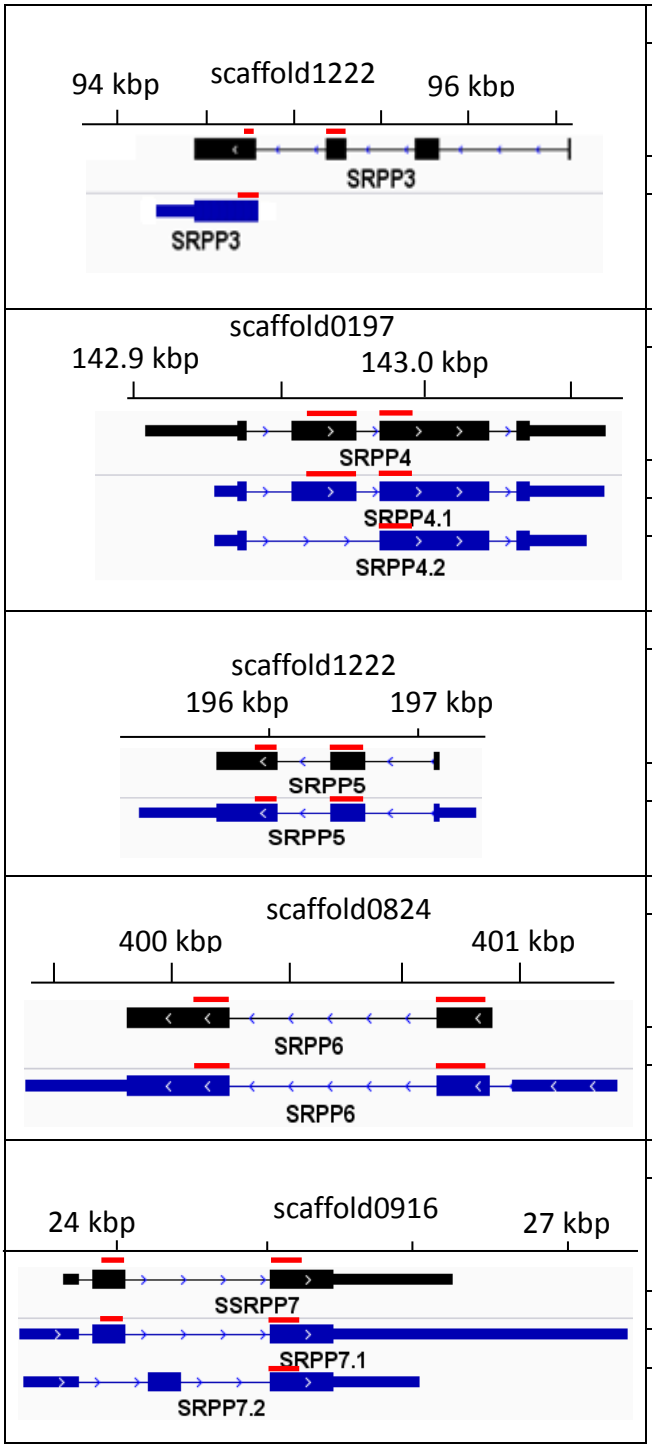


SRPP1		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
SRPP1.1	96	98
SRPP1.2	397,947	346,189
SRPP1.3	927	918



SRPP2		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
SRPP2	23,652	21,862

Figure 6.2.1.2 (continued): The gene models for 18 REFSRPP isoforms, with their corresponding transcript variants. The annotated gene model from genome assembly is represented by black rectangle, while the transcripts variants are represented by blue rectangle. The thicker rectangles represent the protein coding region whilst the thinner rectangles represent the untranslated coding region (UTR). The expression level for each transcript variant was calculated based on the mapped RNA-seq reads generated from the RRIM600 and PB235 latex samples in the present study. The REF domain featured in the gene models and transcript variants is denoted in red line.



SRPP3		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
SRPP3	10	8

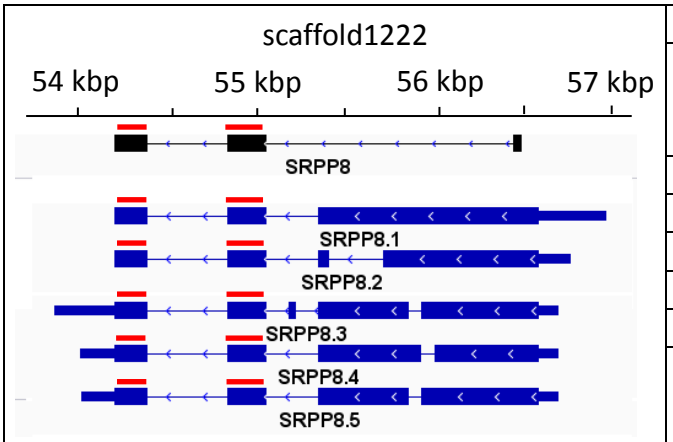
SRPP4		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
SRPP4.1	15	12
SRPP4.2	No expression data	

SRPP5		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
SRPP5	280	1113

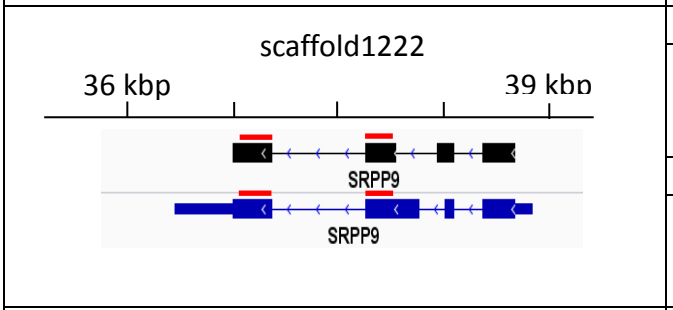
SRPP6		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
SRPP6	60	47

SRPP7		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
SRPP7.1	1,153	1,281
SRPP7.2	60	81

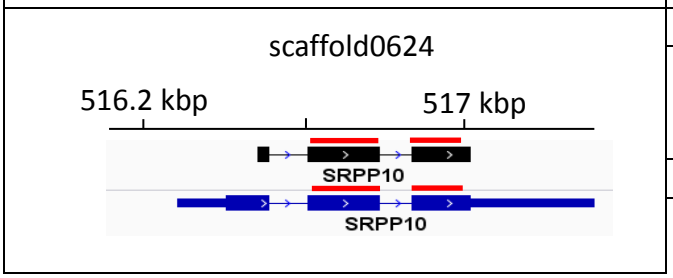
Figure 6.2.1.2 (continued): The gene models for 18 REFSRPP isoforms, with their corresponding transcript variants. The annotated gene model from genome assembly is represented by black rectangle, while the transcripts variants are represented by blue rectangle. The thicker rectangles represent the protein coding region whilst the thinner rectangles represent the untranslated coding region (UTR). The expression level for each transcript variant was calculated based on the mapped RNA-seq reads generated from the RRIM600 and PB235 latex samples in the present study. The REF domain featured in the gene models and transcript variants is denoted in red line.



SRPP8		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
SRPP8.1	111	34
SRPP8.2	No expression	
SRPP8.3	1,669	835
SRPP8.4	No expression	
SRPP8.5	1,484	975



SRPP9		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
SRPP9	520	3,255



SRPP10		
Transcript variants	**Expression level (TPM)**	
	PB235	**RRIM600**
SRPP10	5	6

Figure 6.2.1.2 (continued): The gene models for 18 REFSRPP isoforms, with their corresponding transcript variants. The annotated gene model from genome assembly is represented by black rectangle, while the transcripts variants are represented by blue rectangle. The thicker rectangles represent the protein coding region whilst the thinner rectangles represent the untranslated coding region (UTR). The expression level for each transcript variant was calculated based on the mapped RNA-seq reads generated from the RRIM600 and PB235 latex samples in the present study. The REF domain featured in the gene models and transcript variants is denoted in red line.

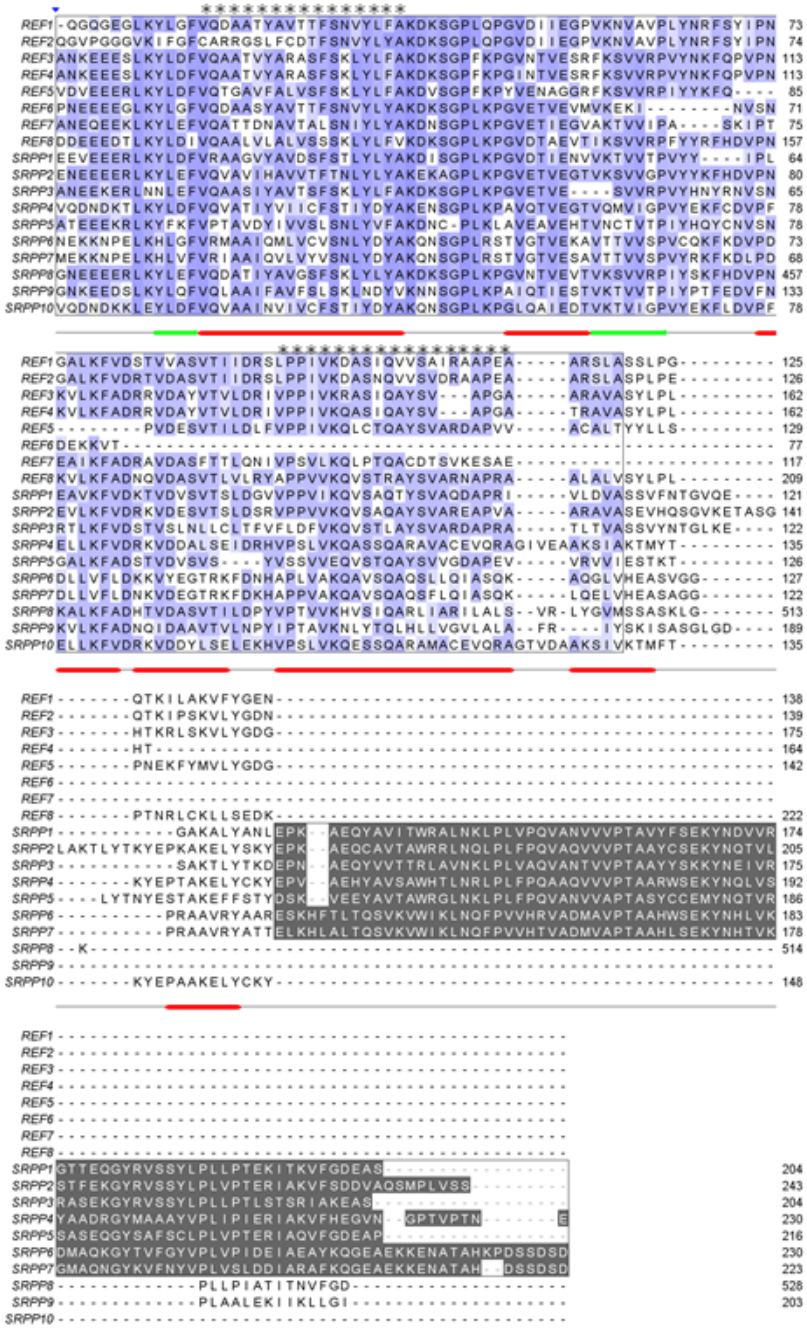


Figure 6.2.1.3: Multiple sequence alignment of part the protein sequence of 18 REFSRPP gene members. The alignment was viewed and edited using Jalview version (Waterhouse et al., 2009). Sequence corresponded to REF domain is shaded in indigo, based on sequence similarity. The darker indigo shade represents higher similarity between the sequences. The region encoding C-terminal of REFSRPP proteins is marked in gray shade. The secondary structure prediction was performed using Jpred 4 (Drozdetskiy et al., 2015). The predicted helices are represented by red tubes while the predicted alpha structure is denoted by green arrows. The transmembrane helices are marked by asterisks (*).

6.2.2. REFSRPP gene family in other plant species

REFSRPP gene orthologs were identified from other plant families (from *Brassicaceae, Euphorbiaceae, Asteraceae*) through reciprocal TBLASTN. These include non-rubber producing plant (*Arabidopsis* and cassava) and non-*Hevea* plants that produce latex (*Taraxacum, Lactuca,* sunflower and *Guayule*). It was observed that REFSRPP genes number varied, ranged from 3 to 10 copies (Table 6.2.2.1). This indicated that while REFSRPP gene number is similar across *Hevea* genotypes, its number differs between different plant species.

Three REFSRPP and six REFSRPP genes were detected from *Arabidopsis* and cassava respectively, with each gene is located on a single chromosome. Multiple copies of REFSRPP genes were detected on the non-*Hevea* rubber-producing plants. In *Lactuca*, 10 REFSRPP genes were detected on three scaffolds, scaffold11, scaffold202 and scaffold203. A cluster of five REFSRPP genes are located on scaffold202, four on scaffold203 and one REFSRPP gene is located on scaffold201. Likewise, multiple numbers of REFSRPP gene were detected from sunflower genome, with four REFSRPP gene are grouped in the same chromosome (chromosome 11). Other individual REFSRPP genes in sunflower were found to be scattered on five different loci (Chromosome 4, Chromosome 8, Chromosome 10, Chromosome 13 and Chromosome 15). The placement of REFSRPP genes in *Taraxacum* and guayule could not be determined due to the lack of the corresponding genomic resources.

Table 6.2.2.1: REFSRPP genes identified from other plant species.

Plant	Annotated gene	Sequence identifier	Protein length	Placement of the gene
Arabidopsis (3 members)	AT1G67360	Ara1	240	Chromosome 1
	AT2G47780	Ara2	235	Chromosome 2
	AT3G05500	Ara3	246	Chromosome 3
Cassava (5 members)	08G117800	Mes1	243	Chromosome 8
	09G170000	Mes2	238	Chromosome 9
	05G063700	Mes3	236	Chromosome 5
	13G012400	Mes4	225	Chromosome 13
	12G011500	Mes5	226	Chromosome 12
Lactuca (10 members)	Lsat_1_v5_gn_9_1021.1	Lsat1	244	Scaffold11
	Lsat_1_v5_gn_9_87781.1	Lsat2	213	Scaffold202
	Lsat_1_v5_gn_9_87820.1	Lsat3	213	
	Lsat_1_v5_gn_9_87841.1	Lsat4	231	
	Lsat_1_v5_gn_9_87901.1	Lsat5	238	
	Lsat_1_v5_gn_9_88040.1	Lsat6	232	
	Lsat_1_v5_gn_0_32660.1	Lsat7	233	Scaffold203
	Lsat_1_v5_gn_9_88301.1	Lsat8	239	
	Lsat_1_v5_gn_9_88241.1	Lsat9	227	
	Lsat_1_v5_gn_9_88221.1	Lsat10	224	
Sunflower (9 members)	HanXRQChr11g0339301	Han1	210	Chromosome11
	HanXRQChr11g0339251	Han2	229	
	HanXRQChr11g0339731	Han3	229	
	HanXRQChr11g0339741	Han4	214	

Table 6.2.2.1 (continued): REFSRPP genes identified from other plant species.

Plant	Annotated gene	Sequence identifier	Protein length	Placement of the gene
	HanXRQChr13g0394341	Han5	247	Chromosome 13
	HanXRQChr15g0465721	Han6	243	Chromosome 15
	HanXRQChr04g0117011	Han7	216	Chromosome 4
	HanXRQChr08g0207791	Han8	519	Chromosome 8
	HanXRQChr10g0284611	Han9	243	Chromosome 10
Taraxacum (possibly 8 members)	DR398691	Tk1	409	No draft genome available
	AGE89406	Tk2	232	
	AGE89407	Tk3	210	
	AGE89408	Tk4	229	
	AGE89409	Tk5	235	
	AGE89410	Tk6	217	
	AMB19721.1	Tk7	210	
	ANP92050.1	Tk8	247	
Guayule (possibly 4 members)	TRINITY_DN8586_c0_g1_i1	Gy1	241	No draft genome available
	TRINITY_DN10263_c1_g1_i1	Gy2	216	
	TRINITY_DN10263_c0_g1_i1	Gy3	166	
	TRINITY_DN1599_c0_g1_i1	Gy4	93	

The observation of REFSRPP gene duplication in rubber producing plants was further confirmed by the constructed phylogenetic tree, using a set of REFSRPP sequences from different species (*Euphorbiaceae, Brassicaceae, Asteraceae, Lauraceae* and monocots). From the phylogenetic tree (Figure 6.2.2.3), the REFSRPP sequences were divided into three main groups. *SRPP4* and *SRPP10* were grouped together with other REFSRPP sequences in Group 1 that were inferred to be involved in the maintenance of membrane structure integrity. Indeed, lipid droplet protein sequences (*LDAP*) which was proposed to be involved in the stabilisation of lipid droplets architecture was also featured in the same group. (Gidda et al., 2016) The expression data for *SRPP4* and *SRPP10* (previously described in Chapter 5, section 5.2.4) showed these genes were lowly expressed in all tissues. It can be deduced that *SRPP4* and *SRPP10* may not be required for the rubber formation in the latex of *Hevea brasiliensis*.

On the other hand, all eleven REFSRPP genes annotated on scaffold1222 were observed to be in Group 2, together with its orthologs from *Brassicaceae*, *Euphorbiaceae* and *Asteraceae*. There was a separation between *Euphorbiaceae* REFSRPP genes with its orthologs from other species. The clade representing the *Euphorbiaceae* showed an inflation of *Hevea* REFSRPP, which was not displayed by cassava, *Jatropha* and *Ricinnus*. In addition, this group featured all REFSRPP genes that were implicated in the rubber formation *in vitro*, namely *REF1* (Dennis and Light, 1989), *SRPP1* (Oh et al, 1999) from *Hevea* and other REFSRPP isoforms from guayule (Kajiura et al., 2018), *Taraxacum* (Laibach et al, 2015) and *Lactuca* (Chakrabarty et al, 2014). Previous rubber elongation in vitro analysis using REF1 and SRPP1 has clearly showed increased uptake of IPP into the rubber molecules. Apart from *REF1*

and *SRPP1, REF3* and *REF7* were also found to be highly expressed in latex tissue (refer to Chapter 5, section 5.2.4). It may indicate that *REF3* and *REF7* were duplicated to supplement the function of *REF1* and *SRPP1* in enhancing rubber formation rate in *Hevea*.

Finally, In Group 3, the majority of the non-*Hevea* REFSRPP sequences in this cluster were functionally annotated as stress-related protein (Kim et al., 2016). *Hevea*'s *SRPP6* and *SRPP7* sequences were found to be clustered in this group. Based on the expression data described in the previous chapter (Chapter 5, section 5.2.4), *SRPP6* and *SRPP7* were lowly expressed in the latex tissue and highly expressed in the root and flower tissues, respectively. Therefore, it may indicate that both *SRPP6* and *SRPP7* are not involved in rubber formation.

The lack of pseudomolecule information for *Hevea* draft genome means that the REFSRPP genes could not be ordered or anchored to the current *Hevea* genome. However, by using the information from a better annotated genome assembly of a closely-related species, information concerning the gene and its pseudomolecule information can be inferred. There was a report demonstrating cassava and *Hevea* genomes were resulted from a paleotetraploidisation event of an ancestral genome (Bredeson et al., 2016, Pootakham et al., 2017). From the phylogenetic tree in Figure 6.2.2.3, REFSRPP gene from cassava (known as *Mes4*) located on Chromosome 9 was found to be the closest in phylogenetic relation to *Hevea* REFSRPP gene members located on scaffold1222. Indeed, a pairwise sequence alignment analysis (data not shown) between Chromosome 9 (from cassava) and scaffold1222 (from *Hevea*) has identified the neighbouring genes flanking

REFSRPP genes on the both syntenic regions. Therefore, pseudomolecule from the cassava genome was inferred to scaffold1222. Figure 6.2.2.4 showed the homologous regions of REFSRPP on Chromosome 9 (from cassava) and scaffold1222 (from *Hevea*). While a single copy of REFSRPP gene (*Mes4*) was retained on Chromosome 9, its homologous region in *Hevea* (scaffold1222) showed a clear expansion of REFSRPP gene. The duplicated REFSRPP genes found in this homologous region are *REF1, REF2, REF3, REF4, REF5, REF7, REF8, SRPP1, SRPP3* and *SRPP8*.

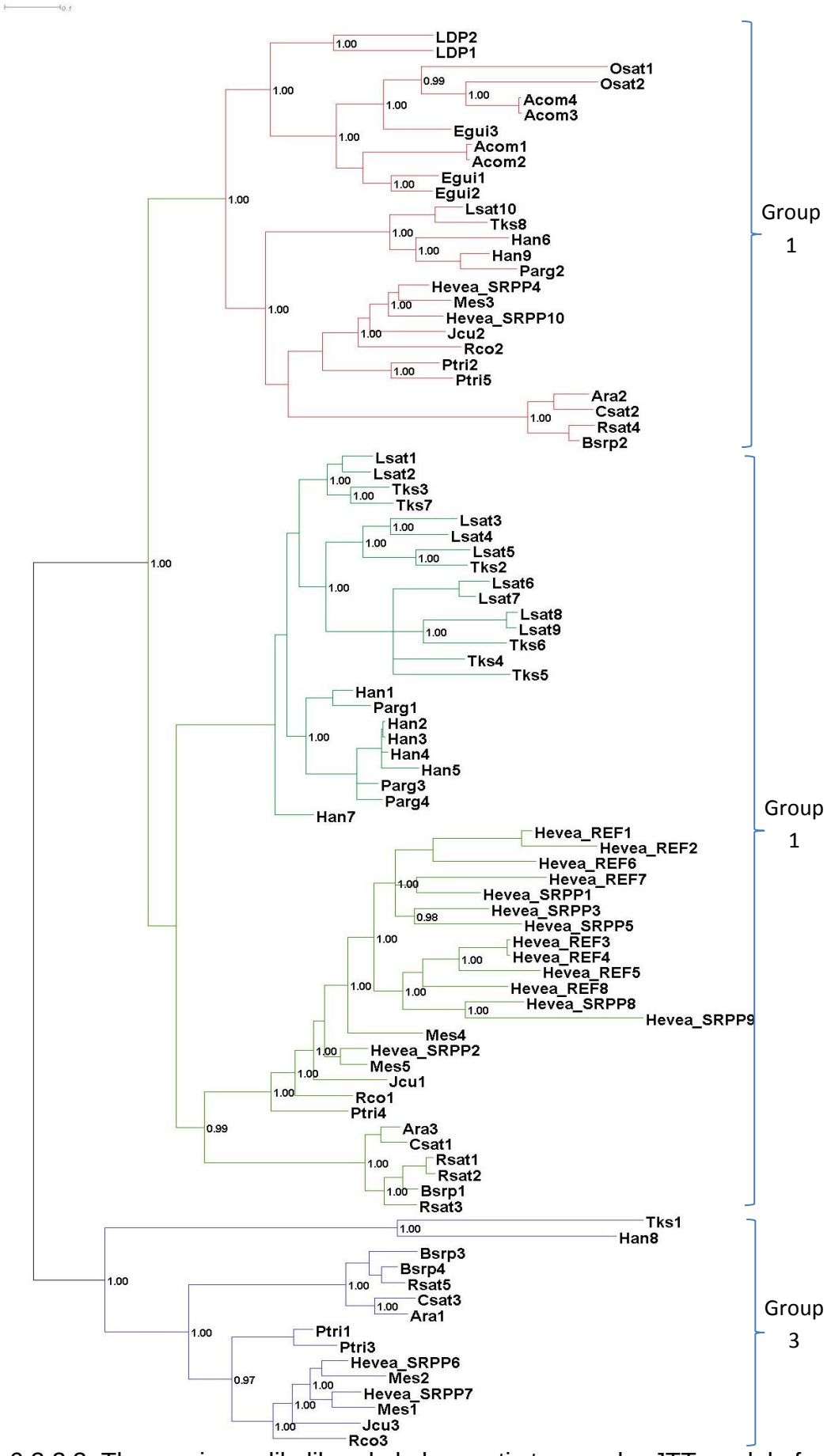


Figure 6.2.2.3. The maximum likelihood phylogenetic tree under JTT model of evolution generated for REFSRPP amino acid sequences from multi-

species plants. The bootstrap (replicates = 1000) analysis was indicated in fraction values located the base of the clades. The sequence identifiers used for REFSRPP orthologs are as follows: Acom: *Ananus comosus*, Ara: *Arabidopsis thaliana*, Bsrp: *Brassica rapa*, Csat: *Camelina sativa*, Egui: *Elaeis guineensis*, Han: *Helianthus annuus*, Jcu: *Jatropha curcas*, LDP: lipid droplet proteins from *Persea americana*, Lsat: *Lactuca sativa*, Mes: Manihot esculenta, Osat: *Oryza sativa*, Parg: *Parthenium argentatum*, Ptr: *Poplar trichocarpa*, Rco: *Ricinnus communis,* Rsat: *Raphanus sativa* and Tks: *Taraxacum koksaghyz*.

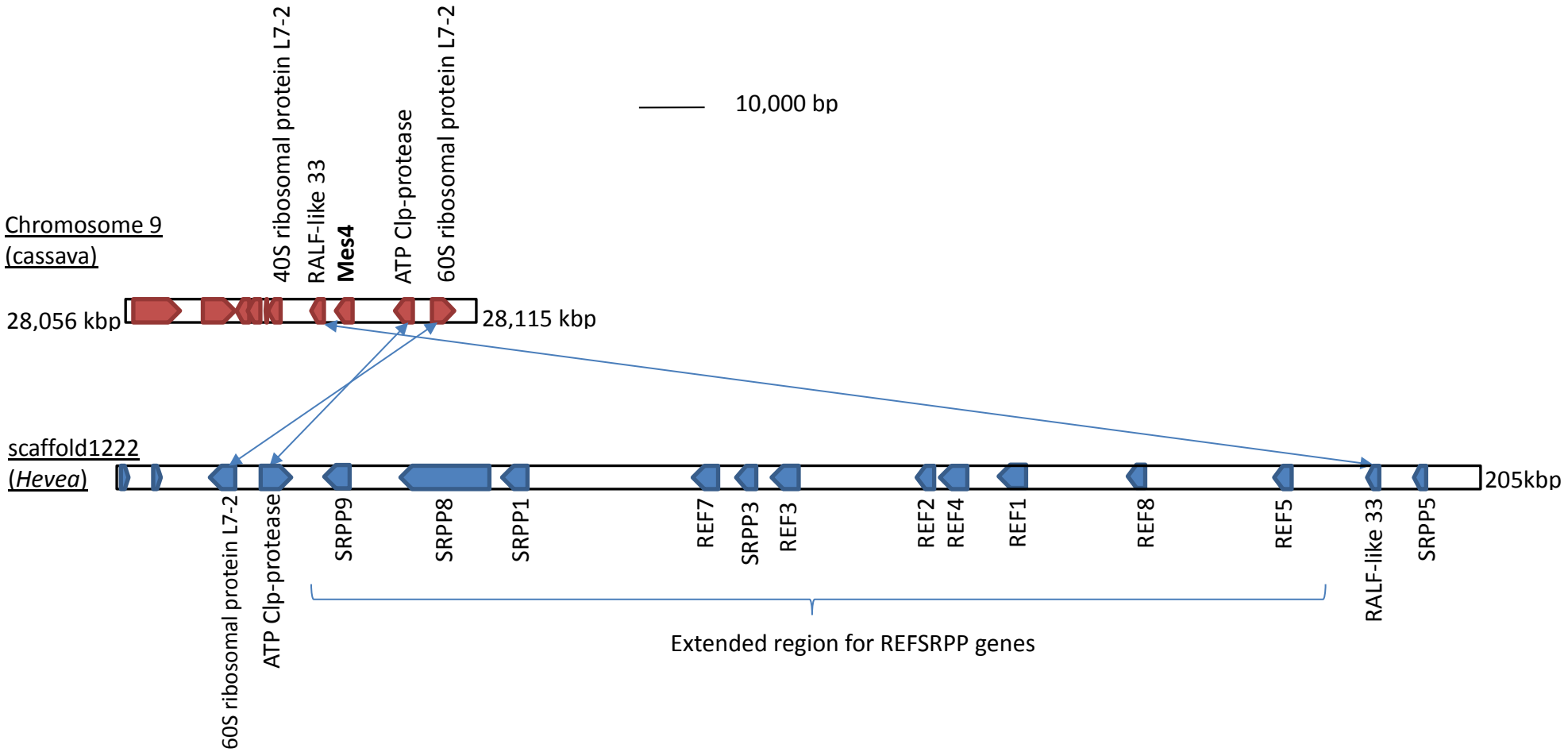


Figure 6.2.2.4: Homologous region of genomic sequences carrying REFSRPP gene from cassava (Chromosome 9; ranged between 28,506 kbp – 28,115 kbp) and scaffold1222 (scaffold length = 205 kbp) from *Hevea*. Based on the previously constructed phylogenetic tree, Mes4 (REFSRPP orthologs) from cassava showed the closest evolutionary relationship. The flanking neighbours of Mes4 (RALF-like 33, ATP p-protease and 60S ribosomal protein L 7-2 genes) are synteny to the corresponding region on *Hevea* draft genome.

6.2.3. Preliminary analysis of the diversity of scaffold1222

Previous results showed that members of the REFSRPP gene cluster on scaffold1222, in particular, REF1 and SRPP1 are involved in the rubber elongation *in vitro*. Thus, it is likely that genes on the scaffold1222 region play an important role in rubber biosynthesis. Therefore, analysis of diversity of the REFSRPP genes on scaffold1222 among *Hevea* genotypes is of interest and may help in explaining variation in latex yield.

Here, a preliminary analysis of scaffold diversity among *Hevea* genotypes was carried out based on the SNP variation. RNA-seq data from 10 *Hevea* genotypes were first used to identify possible polymorphic SNPs. A total of 267 SNP markers were identified using Freebayes (Garrison and Marth, 2012) as described in section 2.34 (Chapter 2) after alignment to the Reyan 7-33-97 draft genome. As the SNPs were identified from the transcribed regions of scaffold1222, any variants located within the non-coding genomic regions or within scaffold gaps would not be detected in this study. Of the 267 SNPs, 140 variants were located on the 5' and 3' untranslated regions (UTR). The remaining SNPs (127 variants) were within the protein-coding regions of the transcribed genes and can be divided into two categories; i) synonymous variations that did not generate amino acid change; ii) non-synonymous variants which led to amino acid changes. In some cases, the variant resulted in a premature stop codon in the coding region. Furthermore, 197 (of the 267) SNP variants were found to be within the REFSRPP genes. The distribution of SNPs on the REFSRPP genes is summarised in Table 6.2.3.1.

Table 6.2.3.1: Distribution of SNPs on the REFSRPP genes

REFSRPP gene	Untranslated regions SNPs			Coding region SNPs		
	3' -end	Intron in between exon	5' -end	Non-synonymous	Synony-mous	Total
REF1	8	2	3	0	1	14
REF2	8	10	4	1	4	27
REF3	9	6	7	0	6	28
REF4	2	5	7	2	6	22
REF5	5	0	4	0	3	12
REF7	2	1	4	0	0	7
REF8	4	6	12	0	4	26
SRPP1	1	0	1	0	2	4
SRPP3	0	1	0	0	0	1
SRPP5	6	5	2	0	1	14
SRPP8	10	12	4	3	5	34
SRPP9	5	1	2	0	0	8

A set of SNPs detected identified in the transcriptome data was developed as Kompetitive allele-specific PCR (KASP) markers (the optimisation protocol and output were described in Chapter 2, section 2.35 and Appendix, section 2.1). Genotyping assays were performed with three SNP markers on 51 *Hevea* genotypes which were chosen to span current *Hevea* genotypes (developed through the Malaysian Rubber Board breeding program) and old genotypes (vegetatively-propagated from the trees germinated from the seeds collected from the Amazonian rainforest). The KASP genotyping results were consistent with the RNA-seq data and known the *Hevea* pedigrees.A simple haplotype were predicted from three SNP markers (Table 6.2.3.2) and its corresponding individual SNP locations are illustrated in Figure 6.2.3.1. A total of 9 simple haplotypes were generated from these three SNP markers. The inheritance of these 9 haplotypes among lines from the Malaysian current pedigree germplasm were compared. Figure 6.2.3.2 illustrates an example of the relationship of 12 (of 51) *Hevea* genotypes based on the heritability of the SNP markers. The results from this pilot analysis indicated a considerable diversity, which is retained even in modern clones.

Table 6.2.3.2: Predicted haplotype from three SNP markers (SNP 1, SNP 2 and SNP 4) located on scaffold1222

	Haplotype combination
Haplotype 1	Parent 1: AAA Parent 2: AAA
Haplotype 2	Parent 1: AAA Parent 2: TAA
Haplotype 3	Parent 1: AAA Parent 2: TAG
Haplotype 4	Parent 1: AGA Parent 2: AAA
Haplotype 5	Parent 1: AGA Parent 2: AGA
Haplotype 6	Parent 1: AGA Parent 2: TAG
Haplotype 7	Parent 1: AGA Parent 2: TGG
Haplotype 8	Parent 1: TAG Parent 2: TGG
Haplotype 9	Parent 1: TGG Parent 2: TGG

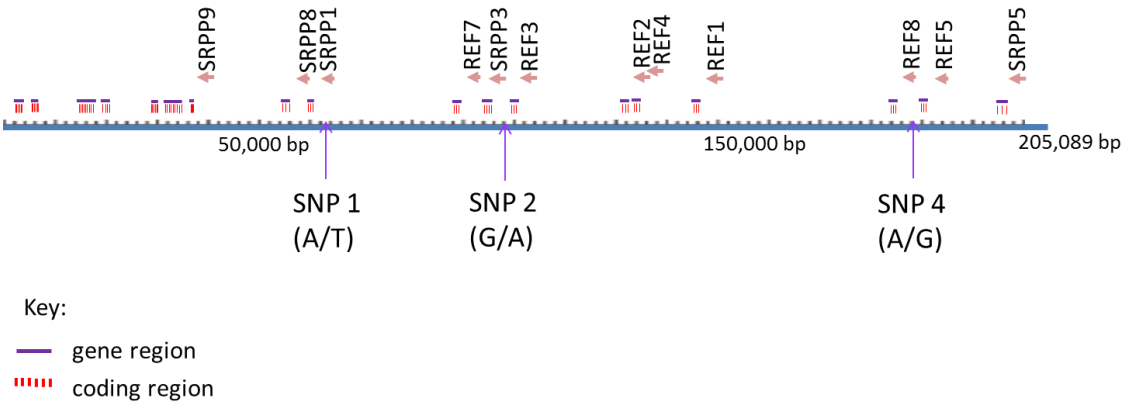


Figure 6.2.3.1: Corresponding SNPs in the KASP genotyping assay, incorporated in the haplotype prediction.

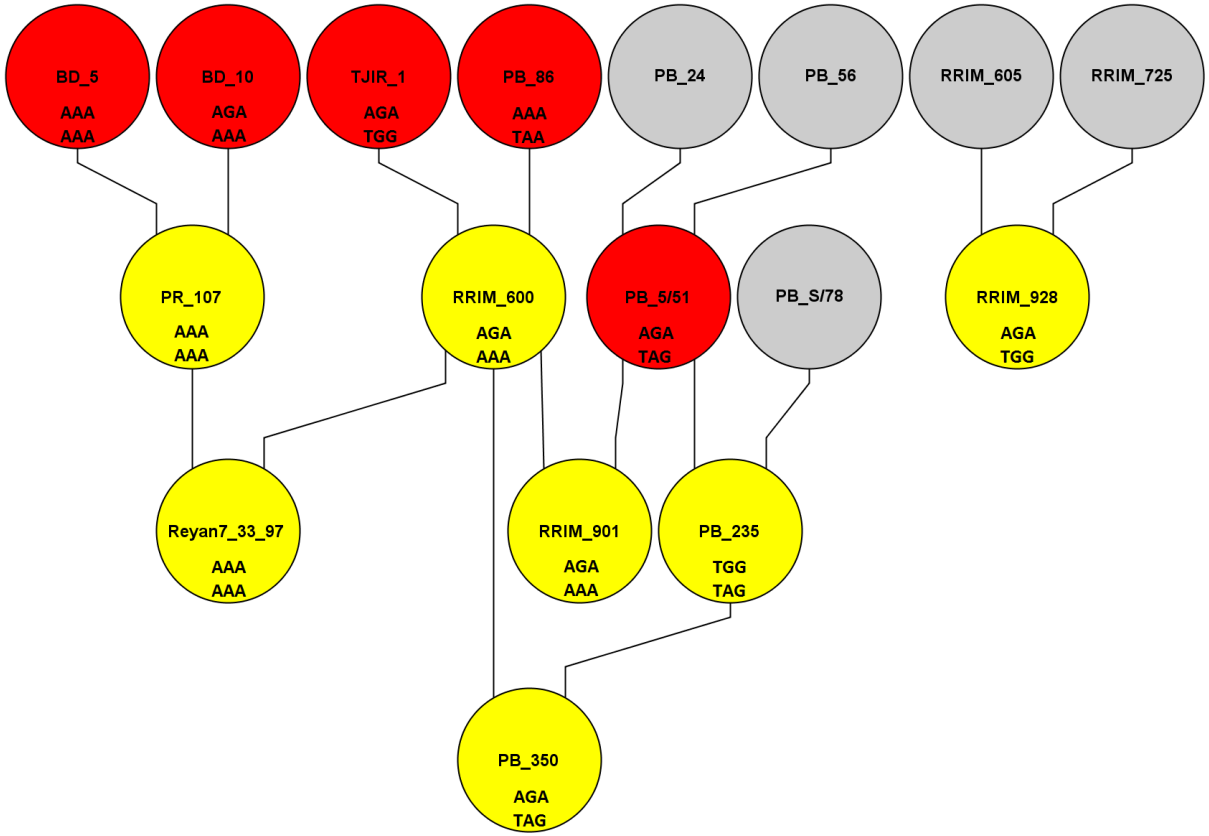


Figure 6.2.3.2: The representation of 12 *Hevea* genotypes, viewed using Helium software (Shaw et al., 2014). The representation of these genotypes was constructed based on the pedigree map information, which was obtained from the Malaysian Rubber Board. The yellow bubbles indicate genotypes of which their alleles were identified by both KASP genotyping and SNP identification through RNA-seq dataset. The red bubbles are the genotypes that have no RNA-seq data but their allelic information was confirmed through KASP genotyping. The grey bubbles indicate parents with unavailable SNP information. The combination of the haplotypes predicted from the three SNP markers used in the genotyping is listed below the genotype identity.

6.3. Discussion

In this study, the annotation of gene models and transcript variants for the 18 REFSRPP gene family members utilising the constructed reference transcriptome were fully described. Of these, *REF6* was found to be truncated, while the *SRPP3* gene model lacked any supporting transcriptomic evidence. Therefore, both might be considered as pseudogenes. In addition, the REF domain first characterised by Sookmark et al (2002) was confirmed as the diagnostic motif for the gene family. The domain was is also detected in several stress proteins and lipid droplet proteins and not restricted to proteins associated with rubber particles. In this study, it was observed that the domain was split by an intron and the intactness of the domain seemed to affect the expression of the transcript variants for each REFSRPP gene. While the REF domain defines REFSRPP gene family as a whole, the REF and SRPP subfamilies are defined by the absence or presence of an additional 3' protein region. These shortened REF proteins are unique to *Hevea* and it is likely that they have evolved by a region loss from an ancestral SRPP sequence. Transmembrane regions are predicted in the REFSRPP sequences and this indicates that these sequences are indeed associated with biological membrane. Previous studies indicated that both *REF1* and *SRPP1* are located on the surface of rubber particles (Yeang et al., 1996, Singh et al., 2003, Berthelot et al., 2012).

The definition system based on the particle localisations, interactions with lipids and epitope was utilised for REFSRPP genes (Dennis and Light, 1989b, Oh et al, 1999, Wadeesirisak et al., 2017). However, these studies were performed using only some of the REFSRPP gene members. When additional REFSRPP gene members were identified, the above nomenclature system is

not really useful to define REF and SRPP sequences. The present study has demonstrated the presence of a longer C-terminal region consistently in SRPP and not in the REF genes. Earlier studies using a limited number of REFSRPP genes also reported the extended C-terminal region in the SRPP proteins (Berthelot et al., 2014a, Tong et al., 2017). The combination of REF domain and the C-terminal region can be therefore, be used for REFSRPP gene classification. Indeed, based on this classification, three of the REFSRPP gene members identified as SRPP (*SRPP8, SRPP9* and *SRPP10*) should be re-classified as REF as they clearly lacked the presence of the extended C-terminal region.

Based on the comparison of expression levels in multiple tissues, most REFSRPP gene members are preferentially expressed in the latex of *Hevea brasiliensis*. Gomez and Moir (1979) reported the present of the laticifer in other *Hevea* tissues such as bark, flower, leaf and root, with traces of latex can be extracted from them. However, REFSRPP genes were not highly expressed in these tissues. Thus, the transcriptional control of these genes may be key to latex rubber biosynthesis. Indeed, the REFSRPP gene expression levels were found to be significantly higher in the high yielding *Hevea* genotypes compare to lowed yielding ones (Venkatachalam et al., 2007, Priya et al., 2007) and the four REFSRPP genes that were highly expressed in latex (*REF1, REF3, REF7* and *SRPP1*) are all on scaffold1222. Therefore, it is likely that the local expansion of this gene family in the scaffold 1222 region is a key event in latex production. Generally, REFSRPP genes were expressed at a higher level in RRIM600. However, one notable observation was for REF8, whereby its dominant transcript (REF8.2) was found to be highly expressed in *Hevea* genotype PB235 compared to RRIM600. REF8 is not included amongst

REFSRPP gene members that involved in rubber elongation activity. Further analysis is warranted to determine whether such expression pattern is persistent among other *Hevea* genotypes. This will help to determine whether this differential expression can be linked to any important traits in *Hevea.*

Bredeson et al. (2014) has reported that the genomes of both *Hevea* and cassava share an ancestral whole genome duplication. Indeed, there is significant colinearity between these two plant species, based on the analysis of conserved linkage blocks Pootakham et al (2017). In the present study, the expansion of REFSRPP genes on scaffold1222 is demonstrated through the comparison homologous genomic regions of cassava and *Hevea*. Although the exact duplication event of REFRPP is still unclear, it can be assumed that after the genome duplication event between cassava and *Hevea* is likely to be due to unequal crossing over giving rise to tandem duplication of REFSRPP genes in *Hevea*.

The duplication of gene can cause functional redundancy, which leads to either the formation of pseudogenes, or sub-functionalisation of the gene members (Flagel and Wendel, 2009). Pseudogenes might be non-functional and often not transcribed or translated into a fully functional protein (Mighell et al., 2000). At the same time, gene duplication can facilitate in the generation of extra transcripts to meet the demand of the associated biological functions (Koonin and Wolf, 2010). It is likely that the expansion of REFSRPP gene family is correlated with the high rate of rubber formation in the *Hevea* laticifers. There also appears to be some evidence of functional diversification of the family over evolutionary time. The REFSRPP genes are distributed across three different clades and each clade appears to have different biological functions.

Previous work has shown that apart from having a role in rubber formation, REFSRPP genes were also implicated in maintaining rubber particle structure integrity as well as in stress-related functions (Dai et al., 2013, Dai et al., 2017, Laibach et al., 2018).

The REFSRPP gene family expansion in *Hevea* also appears to be lineage-specific. All REFSRPP genes located on scaffold1222 were found to be grouped into a single clade in the phylogenetic tree analysis. Comparable gene family expansion also occurs in non-*Euphorbiaceae* rubber producing plants such as lettuce and *Taraxacum*. However, in these cases, their genes are longer (between 210 to 409 amino acids) with no evidence of REF-like truncation. In both lettuce and *Taraxacum* (Reyes-Chin-Wo et al., 2017, Lin et al., 2018) it is SRPP type genes that are correlation to the latex yield. Therefore, it might be possible that the taxonomically distinct expansion of REFSRPP on scaffold1222 and the truncated REF type genes it contains may be key to the ability of *Hevea* to produce large amounts of rubber.

A preliminary study of scaffold 1222 demonstrated indicates that it is polymorphic even in the recent elite Malaysian pedigree lines. Previously, Tang et al (2016) has defined scaffold1222 as a low genetic diversity region, with 14 SNP markers were identified from that genomic sequence. In contrast, this study has identified a total of 197 (of 267) SNP markers to be located on the REFSRPP gene, with 55 non-synonymous SNPs and 32 synonymous SNPs. The 14 SNP markers identified by Tang et al (2016) were also detected in this study. In the future, it might be worthwhile to explore the potential relationship of the region to latex production among *Hevea* lines.

In conclusion, REFSRPP gene family showed a local expansion which appear to be unique to *Hevea.* A pilot study has demonstrated there is considerable diversity of the genomic region containing the duplicated REFSRPP genes. It is clear that a more extensive study of diversity in this region in available germplasm combined with latex yield data is merited. Diversity study of scaffold1222 can be performed through a genome-wide association study (GWAS) or through genetic mapping of a segregation cross involving high and low latex production clones. However, such strategies need a considerable investment to either generate a proper trial of population involving multiple genotypes for GWAS or in preparing *Hevea* mapping population for genetic mapping. In addition, it is imperative that the genome resource is anchored to mapping population and equipped with pseudomolecule information. Such ordered genome resource is important for the analysis of the potential recombinant sites that may occurred during recombination event.

Chapter 7

General discussions

7.1. Conclusions and general discussions

Hevea brasiliensis latex is cytoplasm of laticifer cells and hence contains constituents typical of plant cells (Gomez and Moir, 1979, Gomez and Samsidar, 1989, Gomez, 1990). The unique feature of this exudate from the *Hevea* tree is that it contains a substantial quantity of high molecular weight *cis*-polyisoprene molecules (natural rubber), encapsulated in spherical membrane-bound particles (Ho et al., 1975b). The tapped latex from cultivated *Hevea* is the primary source of natural rubber. Rubber and other non-rubber isoprenoids are synthesised from isoprenyl pyrophosphate (IPP) generated from the cytoplasmic mevalonate (MVA) pathway and the plastidic methyl erythritol phosphate (MEP) pathway (Sando et al., 2008a, Sando et al., 2008b, Chow et al., 2012). As might be expected, the study of the two IPP-producing pathways for isoprenoid biosynthesis pathways, especially rubber formation, is a topic of immense interest in *Hevea* breeding.

Hevea is a perennial crop, and it takes about five years for *Hevea* to reach maturity. A complete breeding cycle for *Hevea* genotype improvement will take between 25 to 30 years to complete (Malaysian Rubber Board, 2009). *Hevea* needs a large planting acreage for commercial cultivation or performance appraisal of a new line. In field conditions, heterogeneity of the samples is common and may affect the interpretation of data. Apart from the heterogeneity factor, the latex samples used in this study tend to naturally coagulate a few hours after exudation from the *Hevea* laticifers. Therefore, appropriate collecting and processing strategies of the samples were practiced in dealing with fresh latex so that the metabolic state of the samples can be preserved. In the present study, the transcript levels of the genes involved in rubber and carotenoid biosynthesis

were analysed together with their corresponding metabolites in latex from two *Hevea* genotypes (RRIM600 and PB235) with contrasting carotenoid levels.

The separation of latex carotenoids from the latex of RRIM600 and PB235 was achieved by using a high-performance liquid chromatography (HPLC) method. The identification of latex carotenoids was based on comparison with carotenoid standards and mass spectrometry (MS/MS) data of the carotenoid ions. The study showed that the latex samples contained four major carotenoids, including β-carotene, α-carotene, lutein and zeaxanthin. β-carotene was found to be the predominant carotenoid and accumulated in the PB235 latex at a higher level compared to that of RRIM600 latex.

As carotenoid formation also involves the precursor IPP, it is of interest to study the flux of isoprenoid intermediates between the branches of rubber and carotenoid biosynthesis. Sub-cellular compartmentalisation of IPP synthesis has been examined in *Hevea* latex and cross-talk is likely to take place between the plastidic MEP and the cytosolic MVA pathways (Archer and Audley, 1987, McMullen and McSweeney, 1966, Sando et al., 2008b, Chow et al., 2012). In this regard, information about the various IPP pools in latex might help in developing strategies to enhance IPP supply directed towards rubber biosynthesis while reducing competition for IPP by other non-rubber isoprenoids. Therefore, measuring isoprenoid intermediates might help in understanding the biosynthesis of rubber and carotenoid.

The profiling of the targeted isoprenoid intermediates (GPP, GGPP, FPP, MEP, DXP, MVA, IPP and DMAPP) was performed using hydrophilic liquid chromatography with mass spectrometry approaches (HILIC-MS/MS). The developed HILIC-MS/MS method was applied in the analysis of isoprenoid intermediates from potato extracts and *Hevea* latex. While the identification of

the targeted isoprenoid in the potato extracts has demonstrated the utility of the developed HILIC-MS/MS method, its applicability in analysing the *Hevea* extract needed further optimisation. This was due to the salt contamination in the latex extracts that caused high backpressure of the HILIC column. Profiling of these isoprenoids in future will need optimisation of the metabolite extraction from latex in order to remove any carry-over contamination.

The measurement of carotenoid compounds and isoprenoid intermediates in this study was performed using two *Hevea* genotypes containing different levels of carotenoids. In future, inclusion of more diverse *Hevea* genotypes with contrasting carotenoid contents should be adopted for a robust correlation of isoprenoid levels and key genes involved in rubber and carotenoid formation. The expression profiling of carotenoid biosynthetic genes from these two *Hevea* genotypes indicated that the expression of a gene encoding phytoene synthase (PSY) was highly expressed in the latex. The PSY isoform 2 gene expression level is higher (>3-fold) in the latex of PB235 than that in RRIM600 and it likely that this may responsible for higher accumulation of β-carotene.

The findings in this study indicated that the accumulation of β-carotene in *Hevea* latex may be caused by the latex-specific expression of a PSY gene. Therefore, it is probable that *PSY* is a suitable choice of indicator gene for the yellow latex trait in *Hevea brasiliensis.* One of future undertakings is to use *PSY* as a candidate gene in determining genetic loci underlying carotenoid accumulation in *Hevea* latex. If *PSY* can be demonstrated to be tightly linked to the loci that regulate carotenoid accumulation, it can be used as a genetic marker to facilitate the selection of the desirable white latex alleles in *Hevea*

breeding program. The possibility of the *PSY* is a gene candidate to explain the genetic basis of carotenoid accumulation has been examined in *Capsicum,* carrots, wheat and cassava (Huh et al., 2001, Ellison et al., 2017, Colasuonno et al., 2017, Udoh et al., 2017). A key challenge with the candidate gene approach is that it is still not known whether the accumulation of the carotenoid is associated to genetic regulation at the structural level (involving the carotenoid structural genes) or at the functional level (involving transcription factor or other regulatory elements) (Thorup et al., 2000, Dhar et al., 2015). Therefore, it is important to understand the regulation of genes involved in carotenoid and rubber formation.

Reference genome of *Hevea* provides an important resource for genetic analysis, but its coverage and quality are limited by the technology available at the time of their construction. The *Hevea* draft genomes currently available contain a considerable portion of truncated gene models. Therefore, in this study, a reference transcriptome was generated to facilitate the analysis of gene expression in the latex of *Hevea brasiliensis*. As demonstrated in the construction of the reference transcriptome in Chapter 5, the incorporation of RNA-seq and Iso-seq data has increased the completeness of the transcriptome data. The reference transcript database was constructed by merging data from three independent sources, which included RNA-seq datasets (52 datasets), Iso-seq data (downloaded from Poothakham et al, 2017) and full-length cDNA sequences (downloaded from Makita et al, 2017). It should be noted that even in species with good pseudomolecule quality genomes, comparable approaches based on a reference transcriptome are now becoming the method of choice for the analyses of gene expression (Zhang et al., 2017a, Brown et al., 2017b).

The constructed reference transcript database in this study is more comprehensive in providing information about transcript variants. The current project focussed on the manual curation of genes associated with five metabolic pathways, that ultimately produce rubber and carotenoids in *Hevea* latex. This resulted in a set of 115 genes, with 151 transcript variants identified from the reference transcriptome. The constructed reference transcripts data set in this study is more comprehensive than previously available in providing information on transcript variants. However, this transcript collection is far from being complete. Due to time constraints, manual inspection and curation were performed on the genes involved in the key pathways in carotenoid and rubber biosynthesis. A complete reference transcriptome would require resources that are beyond the current project. The manual inspection involved the removal of transcript variant artefacts (that could be caused by mis-assembly) and removal of redundant transcripts. The transcriptome comprehensiveness depends on the genes being expressed and captured at the time the cDNA library being constructed. Therefore, the inclusion of additional transcriptome data, such as RNA-seq or Iso-seq from multiple tissues, in varied biological conditions will increase the robustness of the reference transcriptome.

In the intervening time since the reference transcriptome was initially constructed, the pipeline for analysing transcriptomic data and long-read technology has matured considerably. For example, an efficient framework processing and constructing a reference transcriptome is been actively developed. This includes IDP (Au et al., 2013) for error correction, minimap2 (Li, 2018) for long-read alignment and SQANTI (Tardaguila et al., 2017) for combining short-read and long-read transcripts. In addition, the quality and throughput of long-reads provided by Pacific Biosciences

(https://www.pacb.com) and Oxford Nanopore Technologies (https://nanoporetech.com/) have been improved. The long-read sequences will aid in orienting and organising scaffolds so that the contiguity will be improved. Indeed, the utilisation of long-read technology has been reported to improve the assembly of the 17 Gb repeat-rich polyploid wheat genome (Clavijo et al., 2017). *Hevea* genome has been reported to contain between 60-70% of DNA repeat content and it is rather challenging to resolve this region by just relying on the short reads (Lau et al., 2016, Tang et al., 2016, Mollison et al., 2014). Therefore, for future research, improvement of the *Hevea* draft genome can be achieved using long-read technology. Having a near-complete, high quality draft genome will greatly facilitate genomic, transcriptomic and metabolomic analyses of the desired traits in *Hevea.*

The final component of this study focussed on the rubber particle-specific proteins known as rubber elongation factor (REF) and Small Rubber Particle Protein (SRPP), that are thought to be instrumental in producing high-molecular weight rubber (Tanaka et al., 1985, Dennis and Light, 1989, Oh et al, 1999) . REF and SRPP are two different protein classes and collectively form a large gene family, known as the REFSRPP gene family. Though REF and SRPP genes play an important role in rubber synthesis, information on the REFSRPP gene family is still incomplete. The two names that have been used to describe proteins in this family have often been used interchangeably and this is a source of some confusion. Here, it is proposed that the family is referred as REFSRPP and that it can be subdivided into two major subfamilies, REF and SRPP. The key feature which distinguishes REF and SRPP genes is the extended carboxyl terminal region in *SRPP* genes. There is some evidence that members of the REFSRPP gene family (which show latex-specific transcription activity) positively correlated

to latex yield in *Hevea* genotypes (Priya et al., 2007). In this study, specific isoforms of the REFSRPP gene family (*REF1, REF3, REF7* and *SRPP1*) were found to be highly expressed in latex.

The expansion of REFSRPP genes has been demonstrated through the comparison of syntenic regions between *Hevea* and cassava genomes, together with phylogeny tree of the gene family. The genomic region containing a cluster of REF and SRPP genes (scaffold1222) was found to be homologous to Chromosome 9 in cassava. It seems likely that after the genome duplication which predates the separation of cassava and *Hevea* from a single common ancestor, tandem duplication of REF and SRPP genes occurred independently in *Hevea.* The duplicated genes were observed to form a distinct clade in the phylogenetic tree. Furthermore, the REF subfamily as described in the current study appear to have been derived by truncation from an ancestral REFSRPP sequence that corresponds to the *Mes4* sequence (REFSRPP ortholog) in cassava and are all found on a single genomic scaffold1222 in the *Hevea* genome sequence.

Tang et al has identified 14 SNPs on scaffold1222 region and defined the scaffold to be a region of low genetic diversity. In contrast, the present study increased the number of identified variants to 267 SNPs and a preliminary haplotype analysis with a small subset of the SNPs showed a considerable diversity in *Hevea* breeding germplasm. A more comprehensive study using a larger set of SNPs may give a valuable insight into the role of polymorphism in this region in determining latex yield in commercial *Hevea* genotypes. To investigate a link between sequence diversity in the REFSRPP gene family from scaffold1222 to latex yield will require either a creation of one or more mapping populations segregating for latex yield, or the planting of a yield trial with

sufficient numbers of genotypes to carry out a genome-wide association study (GWAS). However, in either case, such experimental set up for *Hevea* will require considerable investment over a long period and a large planting acreage.

Despite the limitations that were experienced in handling field materials with high heterogeneity, this study has provided two accurate analytical methods to examine the levels of isoprenoid metabolites from *Hevea* latex. In addition, the reference transcriptome generated from long and short reads provides a comprehensive resource for transcriptomic profiling analysis in *Hevea.* The findings from the study provide the necessary transcriptomic resource and analytical framework for the measurement of metabolites in latex, for the future undertakings of genetic marker development to accelerate the breeding activity of *Hevea.*

APPENDIX A

Construction of a reference transcriptome for accurate transcript profiling of the *Hevea brasiliensis* latex

1. Library construction and RNA sequencing

A total of six libraries were constructed from total RNAs of each of two rubber tree genotypes, PB235 and RRIM600. The libraries served as a source of cDNA inserts representing the transcriptome of PB235 and RRIM600 for the generation of short read sequences. Subsequently, the short reads would be used as one component in the generation of reference transcriptome and in the analysis of differential gene expression of key genes involved in the isoprenoid biosynthetic pathway. Following the construction of the libraries, validation was performed by (a) determining the average size of inserts through capillary electrophoresis, (b) detecting double-stranded inserts using a fluorometer and (c) amplification of the inserts by qPCR. The validation results are summarised in Table S1.1. The discrepancy observed in the calculated library quantity from qPCR and Qubit might be attributed to the high single-stranded inserts in the libraries. As a final quality control sequencing with Mi-Seq platform was performed. Sequencing on the Mi-Seq platform allowed sampling of the libraries to generate high quality reads and provide a data set for familiarisation in the read assembly and annotation and assessing the quality before committing to higher volume sequencing on the NextSeq

The Mi-Seq-sequenced libraries produced a total of 46,757,134 paired-end sequence reads, each of which was 35-75 bp in length. The summary of quality assessment following the pre-processing of the reads is shown in Table S1.1. The reads flagged as over-expressedfor each library had >97% sequence identity to REFSRPP and Hevein transcripts. REFSRPP genes were reported to

be the most abundant transcript in latex RNA-seq (Chow et al., 2007) and Hevein was reported to be the most abundant soluble protein in latex (Yeang, 1998) therefore, their over-expression is not caused by any error and it is consistent with the biology of latex. No other over-expressed reads matching non-*Hevea* sequences were found, which inferred minimal sequence contamination occurred during the library preparation. A high proportion of the trimmed Mi-Seq reads was mapped to the two *Hevea* draft genomes produced from rubber tree genotypes RRIM 928 (Mollison et al., 2014) and RRIM 600 (Rahman et al., 2013) (Figure S1.1), which were available at the time the libraries were prepared. This indicated that the libraries contained cDNA inserts that can be sequenced into high-quality short reads.

Table S1.1: Summary of assessment of the library qualities using BioAnalyzer, qPCR and flurorometer assays and the following pre-processing of the MiSeq raw reads.

Library	Average size from BioAnalyzer (bp)	Quantity (nM)		MiSeq sequencing data		
		Calculated from qPCR assay	Calculated from fluorometer assay	Raw data	Trimmed data	Overexpressed reads (from trimmed data)
RRIM600.3	417.00	543.5	240.2	2,058,157	2,053,710	REFSRPP Hevein
RRIM600.4	451.02	876.4	356.2	2,382,421	2,377,658	REFSRPP
RRIM600.5	448.80	473.4	99.3	2,638,478	2,634,876	REFSRPP
RRIM600.6	436.00	570.0	212.8	1,890,508	1,886,478	REFSRPP Hevein
RRIM600.8	443.00	859.4	254.3	1,615,821	1,612,506	REFSRPP Hevein
RRIM600.9	369.00	201.6	163.3	2,562,773	2,559,293	REFSRPP
PB235.3	445.00	663.6	333.1	1,707,387	1,703,807	REFSRPP Hevein
PB235.4	443.00	1381.6	276.5	733,186	731,276	REFSRPP
PB235.5	437.00	521.5	268.7	1,943,968	1,939,725	REFSRPP
PB235.6	396.00	724.1	287.9	937,103	935,210	REFSRPP
PB235.7	423.00	294.0	267.7	2,860,714	2,853,975	REFSRPP
PB235.8	569.00	652.7	188.2	2,048,051	2,040,998	REFSRPP

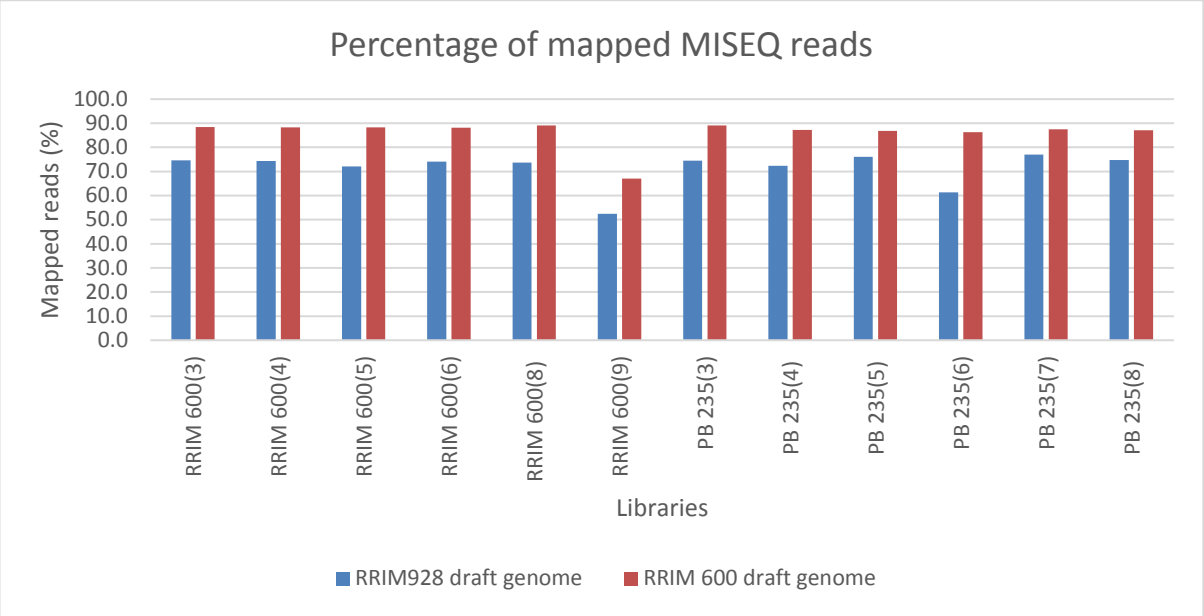


Figure S1.1: Percentage of short reads produced from MISEQ sequencing platform on the inserts of six libraries. The reads were trimmed of sequence adapters and ambiguous data using Trimmomatic software. The mapping of the reads were performed using Tophat from the Tuxedo mapping pipeline (Trapnell et al., 2012).

2. Optimisations of reference-based transcript construction and Iso-seq read error correction

The optimisation pipeline is summarised in Figure S2.1. The pre-requisite steps were performed so that the data will contain a high number of transcripts, with a minimal amount of artefacts (low-quality reads or very low-copy reads).

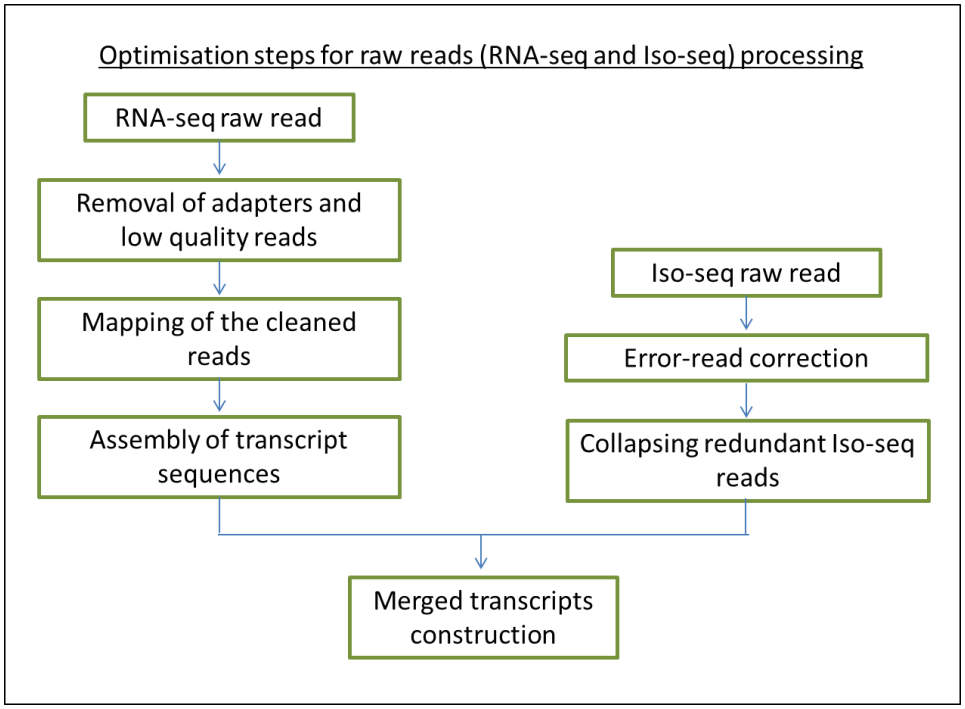


Figure S2.1: The optimisation of raw reads generated from RNA-seq and Iso-seq approaches. Both raw read types needed to be processed prior to the construction of merged transcripts. For RNA-seq, three optimisations were performed, namely finding the best options for i) the removal of adapters and low-quality reads; ii) the alignment of the cleaned reads to the reference draft genome; and iii) the construction of transcript sequences. For the long-read processing, optimisation of the error read correction was performed to remove low-quality bases that gave impact on the transcript's open reading frame. The reference-based assembly of transcripts and collapsed, error-corrected Iso-seq transcripts were subsequently used in the construction of merged transcripts.

2.1. Optimisation of read trimming

The first pre-processing of the reads involved trimming of sequence adapters and removal of low-quality reads from the data. The trimming step was performed using Trimmomatic software (Bolger et al., 2014) on the RNA-seq data generated from one of the total latex RNA libraries (PB235.5). Generally, trimmed or cleaned RNA-seq reads will be free of adapters, and each base exhibits PHRED score (Q) more than 15. The most suitable option that gives the highest number of trimmed reads with minimal loss of data is essential as not only is excessive trimming provide less number of raw reads, it also impacts the downstream analysis of transcriptome profiling (Williams et al., 2016).

In this study, the trimming optimisation was carried out based on the protocol described in Chapter 2 (section 2.21). Two options were evaluated, namely ‘Q threshold’ (Q > 15 and Q > 20) and ‘sliding window’ (window size between 3 nucleotides to 9 nucleotides). During the trimming process, eight different combinations of Q threshold and window size were applied (Figure S2.1.1). The resulted trimmed reads were evaluated based on the number of cleaned reads, the number of reads incorporated in *de novo* transcript construction and Transrate score (Smith-Unna et al., 2016) (Table S2.1.1). The best option for RNA-seq data trimming in this study was determined based on the portion of the trimmed reads incorporated in the *de novo* assembly and Transrate score. From Table S2.1.1, when the trimming was performed using Q threshold > 15 and sliding window=3, most of the cleaned reads were incorporated in the assembled reads. Likewise, the Transrate score indicated that most of the assembled transcripts using these options were supported by RNA-seq reads. Therefore, for the read trimming of all 54 RNA-seq datasets

was performed using Trimmomatic software with Q threshold > 15 and sliding window=3 options.

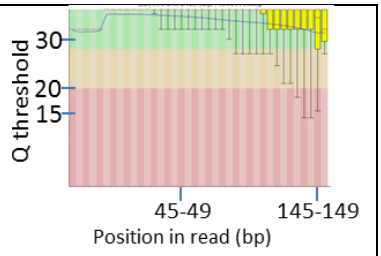
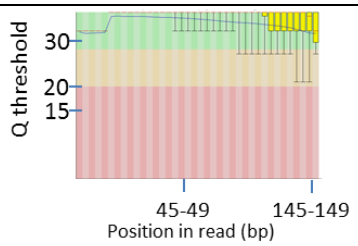
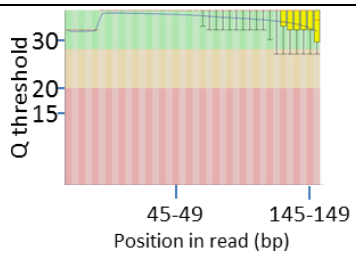
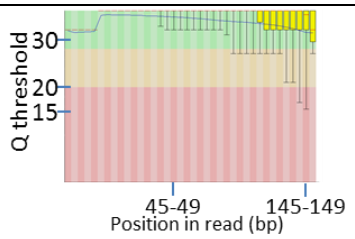
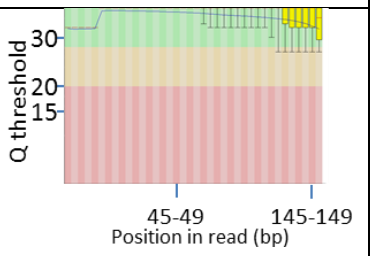
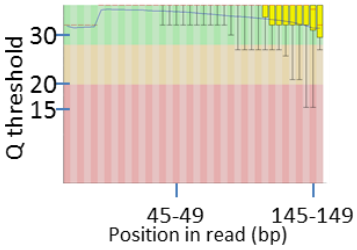
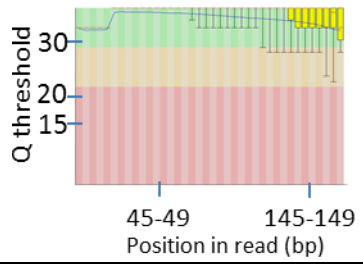
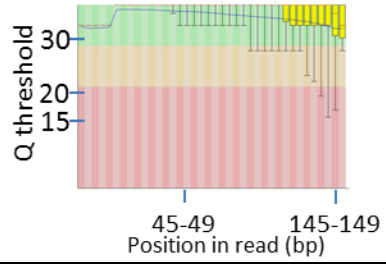
Trimming option 1 Trimming: Q threshold>20 No sliding window		Trimming option 5 Q threshold>15 Sliding window 6	
Trimming option 2 Q threshold>20 Sliding window 3		Trimming option 6 Q threshold>15 Sliding window 7	
Trimming option 3 Q threshold>15 Sliding window 3		Trimming option 7 Q threshold>15 Sliding window 8	
Trimming option 4 Q threshold>15 Sliding window 5		Trimming option 8 Q threshold>15 Sliding window 9	

Figure S2.1.1: Trimming stringencies (options 1-8) executed on one of the latex RNA-seq datasets, summarised as boxplots, generated from FastQC software (Andrews, 2015). The aggressive trimming was in options 1-2 (ranged from strictest to the least stringent), where the filtering was performed without using any sliding window trimmer. The least strict trimming was in options 1, 6, 7 and 8.

Table S2.1.1: The evaluation of trimmed RNA-seq data. The assessment was performed by generating *de novo* transcript assembly using Trinity (Haas, 2016) using k-mer 25, digital normalisation 30. The assembled transcripts that were supported by RNA-seq reads were

evaluated by performing Transrate analysis (Smith-Unna et al., 2016).

	Cleaned pairs	**Number of reads used in the transcript assembly**	**Transrate score**
Latex_RNA-seq (raw)	203,030,672	Not applicable	Not applicable
Latex_RNA trimming option 1	201,806,193	145,478	0.0875
Latex_RNA trimming option 2	179,071,391	164,971	0.1123
Latex_RNA trimming option 3	179,071,391	164,926	0.1924
Latex_RNA trimming option 4	194,543,553	175,728	0.0932
Latex_RNA trimming option 5	197,184,524	178,677	0.0809
Latex_RNA trimming option 6	198,605,526	179,836	0.0848
Latex_RNA trimming option 7	199,162,523	180,637	0.0730
Latex_RNA trimming option 8	199,776,997	181,908	0.0721

2.2. The optimisation of read mapping

Following the determination of the best option for read trimming, the next optimisation step involved mapping of the cleaned reads to a reference sequence (Figure S2.2.1). The primary challenge of this approach was trying to decipher the most optimised position of a large number of short reads across a reference sequence. Additionally, repetitive region in the reference draft genome caused the short reads mapped to multiple places across the reference sequence. Such condition would create ambiguity in the downstream analysis such as expression profiling or polymorphism analysis. Another aspect of RNA-seq mapping is to determine the reads that mapped to splice junctions. A read that aligned to an exon-exon junction has to be spliced across an intron. When a read is split across an intron gap, the aligned segment is referred to as the anchor. In this regard, it is challenging to determine a correct spliced read as the anchor can be as short as one nucleotide. Therefore, it is necessary to use mapping software with the capability to recognise splice junction and map the corresponding reads correctly.

In this study, two splice-aware aligners, STAR (Dobin et al., 2013) and HISAT2 (Kim et al., 2015) were assessed for the reference-based transcript construction. Both aligners used a different algorithm for aligning RNA-seq reads to the reference sequence. Conversely, HISAT aligns RNA-seq read by finding a match for each nucleotide using global index of the reference sequence. When a mismatch is detected (due to read splicing), the correct position of the spliced reads is determined by using the subset of local FM indices. On the other hand, a salient feature of STAR is by aligning RNA-seq reads through long, exact matches using suffix arrays algorithm. When a

mismatch is encountered, the possible positions of the read were found by extending or stitching other unmapped reads.

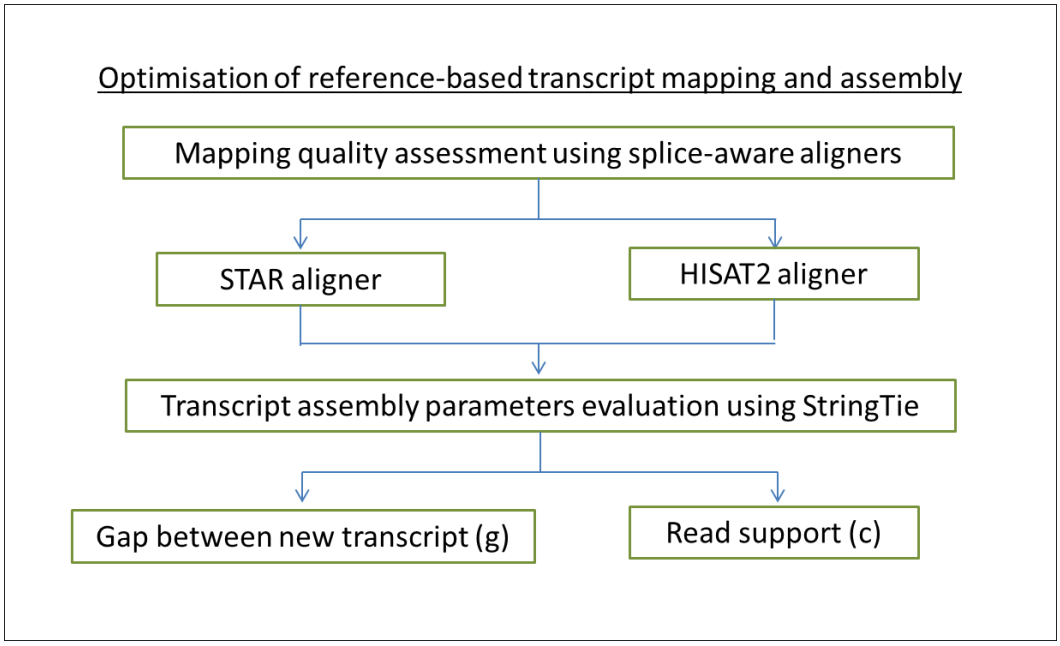


Figure S2.2.1: The optimisation of transcriptome generation from RNA-seq data. The optimisation was performed by evaluating i) mapping quality and ii) assembly options. For mapping quality evaluation, two splice-aware aligners were used, namely STAR (Dobin et al., 2013) and HISAT2 (Kim et al., 2015). When the most suitable aligner was determined, the mapped reads were used for transcript assembly using StringTie (Pertea et al., 2015). During the assembly stage, two options were evaluated as these two options that gave had the most impact on the transcript number and the completeness of transcripts: (i) g, which referred to the minimum locus gap separation value. Reads that were mapped closer than the assessed values were merged in the same processing bundle; and (ii) c, which indicated minimum coverage of reads aligned needed for the predicted transcripts. The completeness of the transcripts was inferred through the percentage of single copy orthologs that were expected to be present across the plant kingdom through BUSCO software.

Both aligners were assessed so that the most suitable mapping software could be used to align a high number of reads accurately, with minimal unmapped reads. The mapping steps using both aligners were performed as described in Chapter 2, section 2.23. The assessment of HISAT2 and STAR was conducted on the alignment of the RNA-seq dataset listed in Table S2.2.1. The raw reads from each dataset were trimmed using Trimmomatic, with optimised options that were determined in the previous section. Besides, three different mismatch rates (0, 2 and 4) were simultaneously compared for each aligner.

Table S2.2.1: RNA-seq datasets used in the assessment of HISAT2 and STAR aligners. The RNA-seq data was downloaded from SRA, NCBI.

Tissue	NCBI accession no.	Read length	Cleaned read pairs
Primary laticifer	SRR5118398	100 x 2	16,836,648
Secondary laticifer	SRR5118397	100 x 2	20,642,818
Root	SRR3136156	100 x 2	25,853,282
Bark	SRR3136158	100 x 2	18,019,834
Leaf	SRR3136159	100 x 2	20,218,651
Latex	SRR3136162	100 x 2	13,224,025
Female flower	SRR3136165	100 x 2	29,411,179
Male flower	SRR3136166	100 x 2	26,242,643
Seed	SRR3136168	100 x 2	24,557,762

The evaluation of mapping quality of RNA-seq reads mapped to the Reyan 7-33-97draft genome was assessed based on the percentage of reads that were uniquely mapped, not-mapped and mapped to multiple-positions in the draft genome (Figure 2.2.1). As expected, the alignments from both HISAT2 and STAR with no mismatch (0 MM) showed the highest portions of unmapped reads, and the proportion decreased for alignments with two mismatches (2 MM) and four mismatches (4 MM). Interestingly, although the possible number of reads to be mapped multiple times was capped at 20, HISAT2 alignments consistently showed a higher portion of multi-mapped reads.

The total number of splice junctions predicted by each aligner is listed in Table S2.2.2. For each aligner, alignment with 0MM showed the lowest number of spliced reads. The number of reads mapped to spliced junctions increased from alignment with 2 MM to alignments with 4MM. The overall pattern showed that HISAT2 aligner consistently predicted a higher number of spliced junctions compared to that of STAR aligner.

Table S2.2.2: Predicted spliced junctions generated from STAR and HISAT2 aligner.

Alignment	Number of splice junctions (> 10 RNA seq read support)
HISAT2_0MM	183,465
HISAT2_2MM	200,852
HISAT2_4MM	211,009
STAR_0MM	154,854
STAR_2MM	183,137
STAR_4MM	189,593

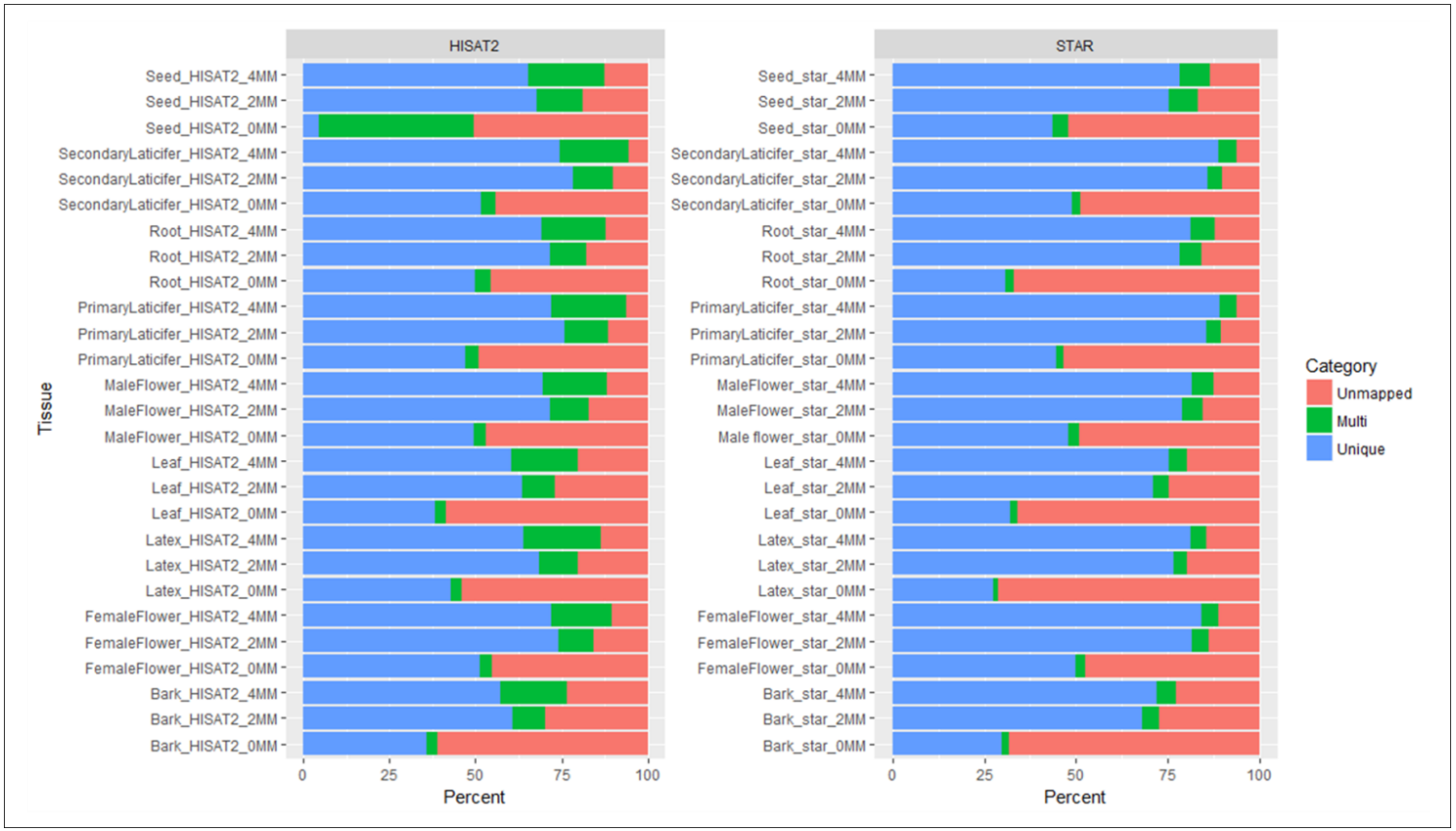


Figure S2.2.2: The evaluation of mapping quality of reads mapped to the *Hevea* draft genome. A total of RNA-seq datasets ranged from different tissues were used in the assessment. Two splicing-aware aligners, HISAT2 (Kim et al, 2016) and STAR (Dobin et al, 2015) were used for the read mapping. Three different mismatch rates were used; no mismatch (0MM), two mismatches per mapped read pair (2MM) and four mismatches per mapped read pair (4MM). “Unique” portion indicates the reads mapped concordantly one time to the reference. “Multi” part represents the reads mapped concordantly to multiple locations (less than 20 multiple locations). “Unmapped” portion indicates read pairs that are not mapped to the reference sequence.

2.3. Optimisation of transcript assembly

A number of transcript assemblers such as StringTie (Pertea et al., 2015), Scripture (Guttman et al., 2010) and Strawberry (Liu and Dickerson, 2017) have been developed to assembled mapped short reads into transcripts. Assembly typically involves the generation of graph structures that represented transcript variants. An accurate transcriptome assembly is an essential pre-requisite for gene expression studies. However, it is highly challenging to infer transcript sequences for very similar gene isoforms accurately. This is because assemblers tend to collapse similar transcript features into a single consensus sequence. Thus, information regarding transcript variants will not be resolved accurately with the collapsed predicted transcripts. Clearly, with the complexity of using short reads in inferring transcript sequences, assessing the quality or completeness of the assembled transcriptome is still a challenge.

In this study, the StringTie assembler (Pertea et al., 2015) was used to construct the aligned RNA-seq into transcript sequences. The software superseded the older assembler version from the same developer, Cufflinks (Trapnell et al., 2012). Apart from reducing computational demands for handling a large amounts of read alignments, this improved software also exhibits superior performance in predicting accurate transcript isoforms (Pertea et al., 2015). Following the spliced-aware alignment of RNA-seq reads, transcript construction was performed using StringTie using the input of mapped reads (0MM and 2MM alignments) from both STAR and HISAT2 aligners. The transcript construction was carried out according to the protocol described in chapter 2 sections 2.3.8.3 and 2.3.8.4. Default option settings were used for the transcript construction, except for read support parameter (known as c; c=2.5 and c=5). This option

specifies the number of minimum coverage of the predicted transcript. Usually, this number can be fractional since some reads can be mapped onto multi places. Conversely, for predicted transcript, reads that aligned in n places onto the transcript will attributed to the 1/n of the coverage. StringTie uses c=2.5 as default.

The completeness of the transcripts was assessed by observing the BUSCO set of single-copy genes expected to be present across plant kingdom through (Simão et al., 2015) (Figure 2.3.1). A transcriptome was considered to have a higher rate of completeness if it covered a greater number of complete single-copy genes in the BUSCO gene database. BUSCO analysis revealed that transcripts were constructed using reads aligned from STAR, consistently gave more complete transcriptome compared to that reads aligned using HISAT2. Additionally, better coverage of BUSCO's single-copy genes was observed when the assembled transcripts were supported by at least five RNA-seq reads. The basic statistics for the two best assemblies from both HISAT2 and STAR (based on BUSCO's evaluation) are summarised in Table S2.3.1. Assemblies of reads aligned using the HISTA2 aligner consistently produced a significantly higher number of transcripts compared to transcripts predicted by STAR-aligned reads. Similarly, transcriptomes that were assembled with two mismatches rate gave higher transcript number compared to that of zero mismatch rate.

Based on optimisation of the mapping of RNA-seq reads and transcriptome assembly from seven datasets, STAR aligner was chosen as the most suitable mapping software to map RNA-seq data to *Hevea* draft sequence. Although HISAT gave a higher number of transcript sequences, STAR has been persistent in generating a higher quality of transcripts (higher number of RNA-seq

mapped onto the reference genome, more complete transcriptome and a lower portion of single-exon transcripts).

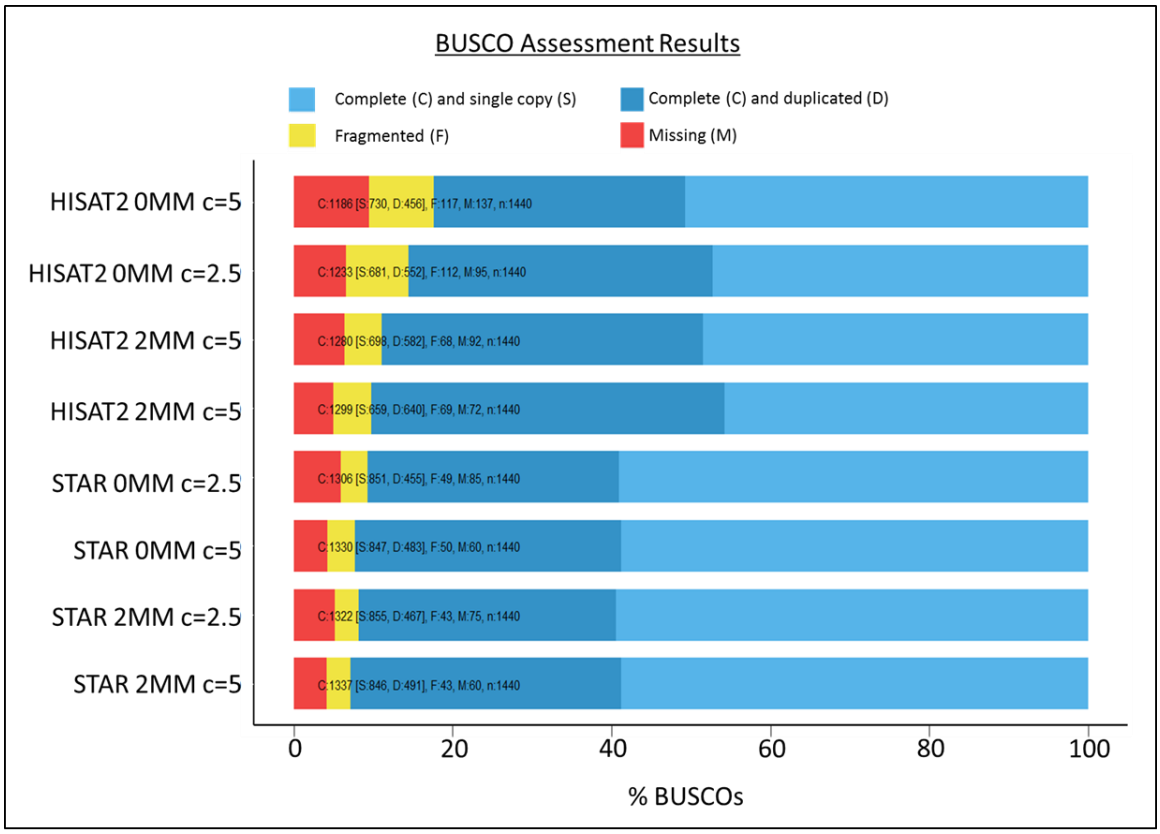


Figure S2.3.1: The completeness of the transcriptome generated from RNA-seq data using HISAT and STAR aligner. During the assembly stage, one option was evaluated, namely read support (c). The option indicated minimum coverage of reads aligned needed for the predicted transcripts. The completeness of the transcripts was inferred through the percentage of single copy orthologs that were expected to be present across the plant kingdom using BUSCO software.

Table S2.3.1: Basic statistics of the assembled transcripts. The statistics were calculated for transcripts that were generated from RNA-seq reads aligned using either HISAT2 or STAR aligners. The reads were aligned with no mismatch (0MM) or with two mismatches rate (2 MM). The assembled transcripts were predicted based on at least support of 5 RNA-seq reads.

Transcriptome	**Predicted transcripts**	**Average length (bp)**	**Longest transcript length (bp)**	**Number of transcripts with a single exon**
HISAT2 0MM	77,743	1,771.7	15,764	12,203 (15.7%)
HISAT2 2MM	80,478	1,912.26	17,793	11,547 (14.3%)
STAR 0MM	72,463	1,764.37	15,755	10,898 (15.0%)
STAR 2MM	73,433	1,967.63	17,222	10,490 (14.2%)

3. Error correction of Iso-seq data

Pootakham et al. (2017) has generated Iso-seq reads from *Hevea* latex to facilitate the annotation of BPM24 genotype draft genome. Additionally, the data (the cleaned circular consensus sequences or CCS) was also utilised in the generation of the reference transcriptome in this study. The inclusion of the Iso-seq reads would enhance the accuracy of isoforms predicted from the previous RNA-seq data. However, despite the ability to generate complete transcriptome, the error rate in Iso-seq data is high (Weirather et al., 2017), with the primary error type is insertion/deletion (indel). If the issue is not addressed, the indels would cause shifts in the transcripts' open reading frame and hence affect the gene model.

To address the impact of indel on the transcripts' coding region, error correction of the Iso-seq data was performed by employing hybrid assembly approach. The hybrid assembly approach utilised RNA-seq short reads to correct indels in the Iso-seq data. RNA-seq short reads were reported to have 0.2% error rate (Abnizova et al., 2017) and the higher accuracy short reads were mapped to the Iso-seq data, to correct the errors. Initially, two algorithm types; LoRDEC (Salmela and Rivals, 2014) and proovread (Hackl et al., 2014) were evaluated for error-correction of the Iso-seq data. Subsequently, the software that produced the best output (based on the assessment of the sequence statistics and BUSCO analysis) were utilised for the generation of merged transcripts.

The basic statistics of transcripts generated from Iso-seq (before and after error correction), are summarised in Table S3.1. A total of 194,135 (proovread) and 190,299 (LoRDEC) of error-corrected Iso-seq sequences were

produced. The average length of the sequences was 1,584 and 1,858 respectively for sequences corrected by proovread and LoRDEC. The average length of the corrected sequences is consistent with being able to span complete transcript sequence, as the average length of predicted reyan transcripts was 1,111 bp. Therefore, Iso-seq could be used to predict full transcripts without the needs to accurately predict the exon/intron boundary. Additionally, the effect of error correction on Iso-seq data was assessed by mapping back the short reads (that were used during hybrid assembly) to the proovread-corrected Iso-seq data. The visualisation of transcripts (before and after error-correction) using Tablet (Milne et al., 2013) (Figure S3.1) showed that proovread had corrected the majority of the indels detected in the raw Iso-seq data. On the other hand, indels were still detected in the Iso-seq data corrected using LoRDEC. Due to proovread generated a higher number of high-quality corrected transcripts, Iso-seq data corrected using the algorithm was chosen to be merged with the assembled RNA-seq transcripts.

Table S3.1: Basic statistics of transcripts that undergone pre- and post-error correction of the Iso-seq data. For comparison, basic statistics for predicted transcripts from *Hevea* draft genome (Tang et al, 2016) and full-length cDNA sequences (Makita et al, 2017) were also assessed.

	Pre-error correction of the Iso-seq data	**Post-error correction of the Iso-seq data**		**CDS_reyan** (Tang et al, 2016)	**RRIM600_flcDNA** (Makita et al, 2017)
		proovread	**LoRDEC**		
Number of sequences	205,310	194,135	190,299	43,877	24,327
Total of bases	353,501,103	307,450,975	353,609,873	48,740,179	16,241,408
Shortest sequence	14	201	200	90	102
Longest sequence	12,659	7,393	12,810	15,948	920
Average	1,722	1,584	1,858	1,111	668
N50	2,109	1,760	2,128	10,236	731
BUSCO assessment[#]	C:56.9%, F:11.9%,M:31.2%	C:72.8%, F:5.8%,M:21.4%	C:72.8%, F:5.8%,M:21.4%	C:90.1%, F:5.4%,M:4.5%	C:7.9%, F:5.9%,M:86.2%

[#] Legend: C: complete single-copy gene; F: fragmented single-copy genes; M: missing single-copy genes
Single-copy genes refer to a set of ortholog sequences that are expected to be present in all plant species (Simão et al., 2015).

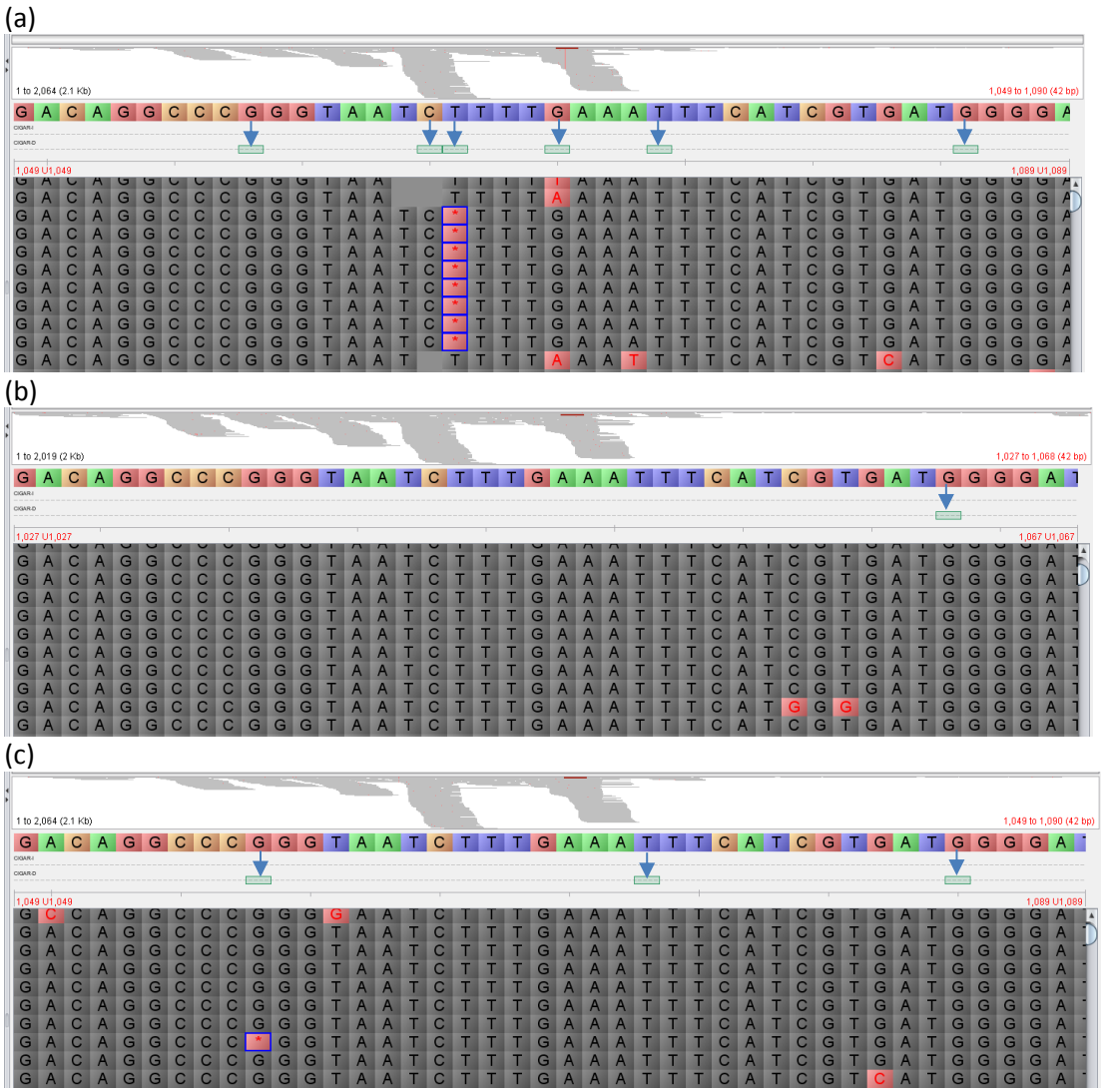


Figure S3.1: View of alignment Iso-seq sequences in Tablet (Milne et al, 2013) for raw Iso-seq data (a); proovread-corrected Iso-seq data (b) and LoRDEC-corrected Iso-seq data. The alignment files were generated by mapping the RNA-seq data (used during error correction in hybrid assembly) to the corresponding Iso-seq data. The error (identified as deletions in this example) correspond to green rectangles and are marked by arrows. Iso-seq data corrected by proovread showed only one deletion sites in the transcript and LoRDEC-corrected data still has three deletion sites in the reading frame.

Table S3.2: Number of isoforms, number of alternatively spliced products for a set of genes chosen for their relevance to rubber and carotenoid biosynthesis. The transcriptome resource predicts a larger number of alternatively spliced products compared to those based on the draft genome.

Pathway	Gene of interest	Isoform predicted by draft genome	Isoform predicted by transcriptome resources	Number of alternatively-spliced products predicted by draft genome	Number of alternatively spliced products predicted by transcriptome resource
MVA	Acetyl-CoA C-acetyltransferase (AACT)	4	4	4	7
	Hydroxymethylglutaryl-coa synthase (HMGS)	2	2	2	2
	Hydroxymethylglutaryl-coa reductase (HMGR)	6	6	6	9
	Mevalonate kinase (MK)	3	3	3	3
	Phosphomevalonate kinase (PMK)	2	2	2	2
	Mevalonate diphosphate decarboxylase (MVD)	2	2	2	2
	Total	19	19	19	25
MEP	1-Deoxy-D-xylulose 5-phosphate synthase (DXS)	10	10	10	11
	1-Deoxy-D-xylulose 5-phosphate reductoisomerase (DXR)	2	2	2	2
	2- C-methyl-D-erythritol 4-phosphate cytidylyltransferase (MEPCT)	2	2	2	2
	4-(Cytidine 5-diphospho)-2- C-methyl-D-erythritol kinase (CDPMEK)	2	2	2	5

	Cyclodiphosphate synthase (MECPS)	2	2	2	2
	4-Hydroxy-3-methylbut-2-enyl-diphosphate synthase (HDS)	2	2	2	3
	4-Hydroxy-3-methylbut-2-enyl diphosphate reductase (HDR)	2	2	2	2
	Total	22	22	22	27
Rubber elongation steps	Isopentenyl diphosphate Δ-isomerase (IPPI)	2	2	2	3
	Geranyl diphosphate synthase (GPPS)	5	5	5	5
	Farnesyl diphosphate synthase (FPPS)	3	3	3	3
	Geranylgeranyl diphosphate synthase	6	6	6	6
	Terpenoid synthases superfamily	1	1	1	1
	Rubber elongation factor (REF)	8	8	8	17
	Small rubber particle protein (SRPP)	10	12	10	19
	cis-prenyl transferase (CPT)	6	6	6	8
	Rubber biosynthesis stimulator (RBSP)	5	5	5	6
	Rubber biosynthesis inhibitor protein (RBIP)	2	2	2	2
	Total	48	50	48	70
Carotenoid	Phytoene synthase (PSY)	4	4	4	6
	Phytoene desaturase (PDS)	1	1	1	1
	Zeta carotene desaturase (ZDS)	3	3	3	3

15-cis-ζ-carotene isomerase (Z-ISO)	1	1	1	1
carotenoid isomerase (CRT-ISO)	2	2	2	4
Lycopene β-cyclase (LCYB)	1	1	1	2
Lycopene epsilon-cyclase (LCYE)	1	1	1	1
β-carotene hydroxylase (BCH)	3	3	3	3
Carotene hydroxylase	2	2	2	4
Zeaxanthin epoxidase	2	2	2	2
violaxanthin de-epoxidase (VDE)	1	1	1	1
neoxanthin synthase (NSY)	1	1	1	1
9-cis-epoxycarotenoid dioxygenase (NCED/MAX)	9	9	9	10
Total	31	31	31	39

APPENDIX B

REF SRPP gene family characterisation

Table S.6.2.2.4: Details of protein sequences used for the construction of phylogenetic tree of REFSRPP gene

Identifier	Species	Family	Annotation	Protein length
*Hevea*_REF1	*Hevea brasiliensis*	Euphorbiaceae	REF1	138
*Hevea*_REF2	*Hevea brasiliensis*	Euphorbiaceae	REF2	139
*Hevea*_REF3	*Hevea brasiliensis*	Euphorbiaceae	REF3	175
*Hevea*_REF4	*Hevea brasiliensis*	Euphorbiaceae	REF4	164
*Hevea*_REF5	*Hevea brasiliensis*	Euphorbiaceae	REF5	154
*Hevea*_REF6	*Hevea brasiliensis*	Euphorbiaceae	REF6	77
*Hevea*_REF7	*Hevea brasiliensis*	Euphorbiaceae	REF7	117
*Hevea*_REF8	*Hevea brasiliensis*	Euphorbiaceae	REF8	222
*Hevea*_SRPP1	*Hevea brasiliensis*	Euphorbiaceae	SRPP1	204
*Hevea*_SRPP2	*Hevea brasiliensis*	Euphorbiaceae	SRPP2	243
*Hevea*_SRPP3	*Hevea brasiliensis*	Euphorbiaceae	SRPP3	204
*Hevea*_SRPP4	*Hevea brasiliensis*	Euphorbiaceae	SRPP4	230
*Hevea*_SRPP5	*Hevea brasiliensis*	Euphorbiaceae	SRPP5	216
*Hevea*_SRPP6	*Hevea brasiliensis*	Euphorbiaceae	SRPP6	230
*Hevea*_SRPP7	*Hevea brasiliensis*	Euphorbiaceae	SRPP7	223
*Hevea*_SRPP8	*Hevea brasiliensis*	Euphorbiaceae	SRPP8	153
*Hevea*_SRPP9	*Hevea brasiliensis*	Euphorbiaceae	SRPP9	203
*Hevea*_SRPP10	*Hevea brasiliensis*	Euphorbiaceae	SRPP10	148
Ara1	*Arabidopsis thaliana*	Brassicaceae	AT1G67360	240
Ara2	*Arabidopsis thaliana*	Brassicaceae	AT2G47780	235
Ara3	*Arabidopsis thaliana*	Brassicaceae	AT3G05500	246
Mes1	*Manihot esculenta*	Euphorbiaceae	Manes.13G012400.1	261
Mes2	*Manihot esculenta*	Euphorbiaceae	Manes.12G011500.1	226
Mes3	*Manihot esculenta*	Euphorbiaceae	Manes.05G063700.1	236

Mes4	*Manihot esculenta*	Euphorbiaceae	Manes.09G170000.1	238
Mes5	*Manihot esculenta*	Euphorbiaceae	Manes.08G117800.1	243
Ptri1	*Populus trichocarpa*	Salicaceae	Potri.001G055300.1	229
Ptri2	*Populus trichocarpa*	Salicaceae	Potri.002G206000.1	234
Ptri3	*Populus trichocarpa*	Salicaceae	Potri.003G173100.1	231
Ptri4	*Populus trichocarpa*	Salicaceae	Potri.005G025700.1	242
Ptri5	*Populus trichocarpa*	Salicaceae	Potri.014G131100.1	102
Lsat1	*Lactuca sativa*	Asteraceae	LOC111884539	212
Lsat2	*Lactuca sativa*	Asteraceae	LOC111884522	212
Lsat3	*Lactuca sativa*	Asteraceae	LOC111884549	230
Lsat4	*Lactuca sativa*	Asteraceae	LOC111884546	237
Lsat5	*Lactuca sativa*	Asteraceae	LOC111884541	231
Lsat6	*Lactuca sativa*	Asteraceae	LSAT_0X32660	232
Lsat7	*Lactuca sativa*	Asteraceae	LOC111895550	238
Lsat8	*Lactuca sativa*	Asteraceae	LOC111895554	226
Lsat9	*Lactuca sativa*	Asteraceae	LOC111895551	223
Lsat10	*Lactuca sativa*	Asteraceae	LOC111920526	243
Han1	*Helianthus annuus*	Asteraceae	LOC110890441	210
Han2	*Helianthus annuus*	Asteraceae	LOC110890438	229
Han3	*Helianthus annuus*	Asteraceae	LOC110890483	229
Han4	*Helianthus annuus*	Asteraceae	LOC110890486	214
Han5	*Helianthus annuus*	Asteraceae	HannXRQ_Chr13g0394341	147
Han6	*Helianthus annuus*	Asteraceae	LOC110910120	243
Han7	*Helianthus annuus*	Asteraceae	LOC110937262	216
Han8	*Helianthus annuus*	Asteraceae	LOC110871612	519
Han9	*Helianthus annuus*	Asteraceae	LOC110884260	243
Tks1	*Taraxacum brevicorniculatum*	Asteraceae	MF471639.1	409
Tks2	*Taraxacum brevicorniculatum*	Asteraceae	JQ991928.1	232

Tks3	*Taraxacum brevicorniculatum*	Asteraceae	JQ991929.1	210
Tks4	*Taraxacum brevicorniculatum*	Asteraceae	JQ991930.1	229
Tks5	*Taraxacum brevicorniculatum*	Asteraceae	JQ991931.1	235
Tks6	*Taraxacum brevicorniculatum*	Asteraceae	JQ991932.1	217
Tks7	*Taraxacum brevicorniculatum*	Asteraceae	KT899432.1	210
Tks8	*Taraxacum brevicorniculatum*	Asteraceae	KU496880.1	247
Parg1	*Parthenium argentatum*	Asteraceae	Downloaded RNA-seq data	216
Parg2	*Parthenium argentatum*	Asteraceae	Downloaded RNA-seq data	241
Parg3	*Parthenium argentatum*	Asteraceae	Downloaded RNA-seq data	93
Parg4	*Parthenium argentatum*	Asteraceae	Downloaded RNA-seq data	166
Rsat1	*Raphanus sativus*	Brassicaceae	LOC108812405	246
Rsat2	*Raphanus sativus*	Brassicaceae	LOC108811305	246
Rsat3	*Raphanus sativus*	Brassicaceae	LOC108857502	245
Rsat4	*Raphanus sativus*	Brassicaceae	LOC108813444	224
Rsat5	*Raphanus sativus*	Brassicaceae	LOC108843369	240
Bsrp1	*Brassica rapa*	Brassicaceae	LOC103848941	246
Bsrp2	*Brassica rapa*	Brassicaceae	LOC103866345	224
Bsrp3	*Brassica rapa*	Brassicaceae	LOC103831029	232
Bsrp4	*Brassica rapa*	Brassicaceae	LOC103852398	236
Csat1	*Camelina sativa*	Brassicaceae	LOC104712806	246
Csat2	*Camelina sativa*	Brassicaceae	LOC104783366	238
Csat3	*Camelina sativa*	Brassicaceae	LOC104701343	245
Jcu1	*Jatropha curcas*	Euphorbiaceae	LOC105647503	240
Jcu2	*Jatropha curcas*	Euphorbiaceae	LOC105642500	236
Jcu3	*Jatropha curcas*	Euphorbiaceae	LOC105644810	228
Rco1	*Ricinnus communis*	Euphorbiaceae	LOC8287083	241
Rco2	*Ricinnus communis*	Euphorbiaceae	LOC8271757	236
Rco3	*Ricinnus communis*	Euphorbiaceae	LOC8286799	229

Egui1	*Elaeis guineensis*	Arecaceae	LOC105036311	243
Egui2	*Elaeis guineensis*	Arecaceae	LOC105045283	252
Egui3	*Elaeis guineensis*	Arecaceae	LOC105052510	221
LDP1	*Persea americana*	Lauraceae	KF031141.1	250
LDP2	*Persea americana*	Lauraceae	KF031142.1	251
Acom1	*Ananus comosus*	Bromeliaceae	ACMD2_18209	269
Acom2	*Ananus comosus*	Bromeliaceae	ACMD2_18213	269
Acom3	*Ananus comosus*	Bromeliaceae	ACMD2_22278	229
Acom4	*Ananus comosus*	Bromeliaceae	ACMD2_27137	223
Osat1	*Oryza sativa*	Poaceae	Os07g0671800	255
Osat2	*Oryza sativa*	Poaceae	Os05g0151300	253

4. KASP assay

Five primers were selected for *Hevea* tree genotyping assay. The primers were designed to span about 201-bp region bearing a single nucleotide polymorphism (SNP) marker. A total of five primers corresponding to five markers located on scaffold1222 on *Hevea* genome were designed. Subsequently, the markers were used for genotyping of *Hevea* germplasm through Kompetitive Allele Specific PCR (KASP) assay. The results from KASP genotyping assay would provide an experimental evidence that the SNPs identified through RNA-seq were accurate rather than due to mapping errors. The KASP assay was used to genotype 51 *Hevea* lines that originated from the current Malaysian pedigree and few international lines. A summary of the rubber genotypes, their genotype identifiers and country of origin is presented in Table S4.1. The location of SNPs used in KASP genotyping is summarised in Table S4.2.

From the five KASP primers, three (SNP001, SNP002, SNP004) were found to show good clustering of homologous and heterozygous alleles while SNP003 and SNP005 did not show good allelic discrimination (Figure S4.1 a and b). Therefore, the genotyping assay was performed using SNP001, SNP002 and SNP004 markers in order to identify the allelic information of 51 rubber tree genotypes. The genotyping results are summarised in Table S4.3 (a-c). The genotyping assay showed for each SNP marker, the portion of undetermined rubber tree genotypes was between 6 – 10%.

Table S4.1: The list of rubber tree genotypes in the KASP genotyping assay. For the KASP genotyping assay, minimal 40 ng of leaf DNA of an individual rubber tree genotypes were amplified.

Genotype	Parent 1	Parent 2	Origin
1. IAN873	PB86	F1717	Brazil
2. Reyan 7-33-97	RRIM600	PR107	China
3. GT1	Not available		Indonesia
4. PR107	Not available		Indonesia
5. BD10	Not available		Indonesia
6. BD5	Not available		Indonesia
7. TJIR1	Not available		Indonesia
8. PB5/51	PB24	PB56	Malaysia
9. PB235	PB5/51	PBS/78	Malaysia
10. PB260	PB5/51	PB49	Malaysia
11. PB350	RRIM600	PB235	Malaysia
12. PB355	PB235	PR107	Malaysia
13. PB49	Not available		Malaysia
14. PB86	Not available		Malaysia
15. RRIM600	TJIR1	PB86	Malaysia
16. RRIM623	PB49	PILB84	Malaysia
17. RRIM901	PB5/51	RRIM600	Malaysia
18. RRIM908	PB5/51	RRIM623	Malaysia
19. RRIM911	PB5/51	RRIM623	Malaysia
20. RRIM921	PB5/51	FORD351	Malaysia
21. RRIM924	RRIM600	PB5/51	Malaysia
22. RRIM926	RRIM623	PB5/51	Malaysia
23. RRIM927	RRIM600	PB5/51	Malaysia
24. RRIM928	RRIM605	RRIM725	Malaysia
25. RRIM930	RRIM623	RRIM600	Malaysia
26. RRIM937	RRIM703	PB5/51	Malaysia
27. RRIM940	RRIM701	PB5/51	Malaysia
28. RRIM2001	RRIM600	PB260	Malaysia
29. RRIM2002	PB5/51	FORD351	Malaysia
30. RRIM2003	PB5/51	RRIM703	Malaysia
31. RRIM2004	PB5/51	RRIM703	Malaysia
32. RRIM2005	PB5/51	RRIM703	Malaysia
33. RRIM2006	PB5/51	PR261	Malaysia
34. RRIM2007	GT1	PB260	Malaysia
35. RRIM2009	GT1	PB260	Malaysia
36. RRIM2010	RRIM623	PR261	Malaysia
37. RRIM2011	GT1	PB260	Malaysia
38. RRIM2012	GT1	PB260	Malaysia
39. RRIM2013	RRIM600	PR261	Malaysia
40. RRIM2014	RRIM717	PR261	Malaysia
41. RRIM2015	PB5/51	IAN873	Malaysia
42. RRIM2016	PB5/51	IAN873	Malaysia
43. RRIM2017	PB5/51	RRIM623	Malaysia
44. RRIM2018	PB5/51	RRIM623	Malaysia

45.	RRIM2019	PB5/51	IAN873	Malaysia
46.	RRIM2020	PB5/51	IAN873	Malaysia
47.	RRIM2021	PB5/51	IAN873	Malaysia
48.	RRIM2023	IAN873	PB260	Malaysia
49.	RRIM2024	IAN873	PB235	Malaysia
50.	RRIM2025	IAN873	RRIM803	Malaysia
51.	RRIM3001	IAN873	PB235	Malaysia

Table S4.2: Basic summaries of SNPs used in the KASP genotyping assay

SNP identifier	**Allele**		**Position (bp)**	**Localisation**
	Reference	**Alternative**		
SNP001	A	T	60,322	5’ -UTR
SNP002	A	G	95,172	Exon
SNP003	C	T	137,413	Exon
SNP004	A	G	175,369	Exon
SNP005	A	G	181,449	Exon

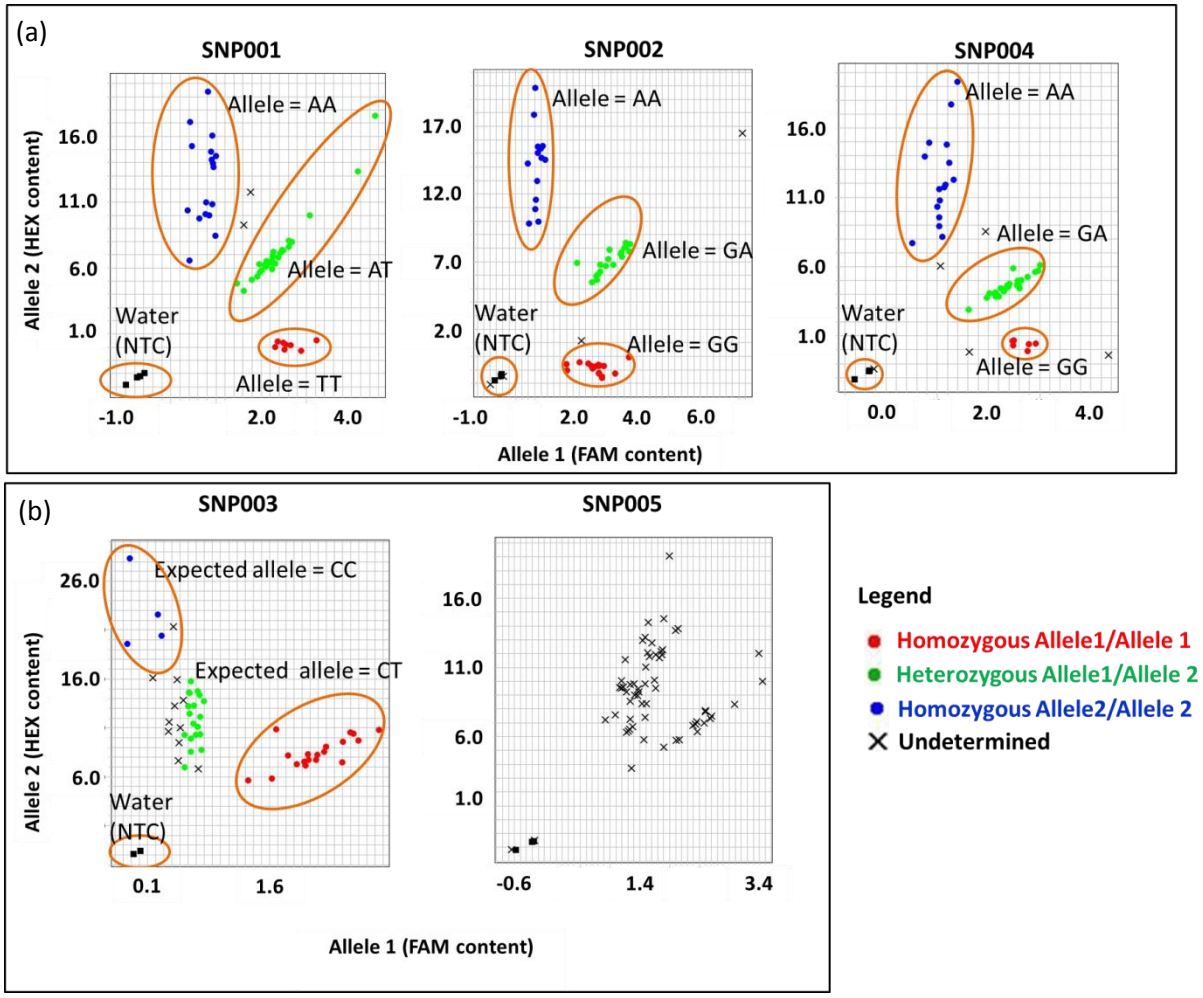


Figure S4.1: Discrimination of heterozygous and homozygous allele genotypes. (a) SNPs that successfully discriminate rubber tree genotypes according to their allele content. (b) SNP markers that gave ambiguous genotyping results, due to the less defined clustering of homozygous and heterozygous alleles.

Table S4.3: Allelic information of 51 *Hevea* genotypes generated through KASP genotyping

(a) SNP001

Clone	SNP call	
RRIM623	TT	8 clones
RRIM908	TT	homozygous
RRIM926	TT	allele 1
RRIM2004	TT	
RRIM2007	TT	
RRIM2014	TT	
RRIM2018	TT	
PB235	TT	
RRIM600	AA	14 clones
BD5	AA	homozygous
IAN873	AA	allele 2
RRIM901	AA	
RRIM927	AA	
RRIM937	AA	
RRIM2002	AA	
RRIM2005	AA	
RRIM2006	AA	
RRIM2013	AA	
RRIM2016	AA	
RRIM2021	AA	
RRIM2023	AA	
PR107	AA	
RRIM928	AT	26 clones
GT1	AT	heterozygous
PB5/51	AT	
PB49	AT	
PB86	AT	
PB350	AT	
PB355	AT	
TJIR1	AT	
RRIM911	AT	
RRIM921	AT	
RRIM924	AT	
RRIM930	AT	
RRIM940	AT	
RRIM2001	AT	
RRIM2003	AT	
RRIM2009	AT	
RRIM2010	AT	
RRIM2011	AT	
RRIM2012	AT	
RRIM2015	AT	
RRIM2017	AT	
RRIM2019	AT	
RRIM2020	AT	
RRIM2024	AT	
RRIM2025	AT	
RRIM3001	AT	
BD10	Undetermined	
PB260	Undetermined	
REYAN	Undetermined	

(b) SNP002

Clone	SNP call	
RRIM928	GG	15 clones
PB49	GG	homozygous
TJIR1	GG	allele 1
RRIM623	GG	
RRIM911	GG	
RRIM930	GG	
RRIM2002	GG	
RRIM2006	GG	
RRIM2009	GG	
RRIM2010	GG	
RRIM2014	GG	
RRIM2016	GG	
RRIM2017	GG	
RRIM2023	GG	
RRIM2024	GG	
RRIM600	AA	13 clones
BD5	AA	homozygous
PB86	AA	allele 2
PB355	AA	
RRIM924	AA	
RRIM2001	AA	
RRIM2003	AA	
RRIM2011	AA	
RRIM2012	AA	
RRIM2013	AA	
RRIM2019	AA	
REYAN	AA	
PR107	AA	
BD10	GA	19 clones
GT1	GA	heterozygous
IAN873	GA	
PB5/51	GA	
PB350	GA	
RRIM901	GA	
RRIM921	GA	
RRIM926	GA	
RRIM927	GA	
RRIM940	GA	
RRIM2004	GA	
RRIM2005	GA	
RRIM2007	GA	
RRIM2015	GA	
RRIM2018	GA	
RRIM2020	GA	
RRIM2025	GA	
RRIM3001	GA	
PB235	GA	
PB260	Undetermined	
RRIM908	Undetermined	
RRIM937	Undetermined	
RRIM2021	Undetermined	

(c) SNP002

Clone	SNP call	
RRIM623	GG	6 clones
RRIM926	GG	homozygous
RRIM2004	GG	allele 1
RRIM2007	GG	
RRIM2018	GG	
PB235	GG	
RRIM600	AA	15 clones
BD5	AA	homozygous
IAN873	AA	allele 2
PB86	AA	
RRIM901	AA	
RRIM927	AA	
RRIM937	AA	
RRIM2002	AA	
RRIM2005	AA	
RRIM2006	AA	
RRIM2013	AA	
RRIM2016	AA	
RRIM2023	AA	
REYAN	AA	
PR107	AA	
RRIM928	AG	25 clones
GT1	AG	heterozygous
PB5/51	AG	
PB49	AG	
PB350	AG	
PB355	AG	
TJIR1	AG	
RRIM911	AG	
RRIM921	AG	
RRIM924	AG	
RRIM930	AG	
RRIM940	AG	
RRIM2001	AG	
RRIM2003	AG	
RRIM2009	AG	
RRIM2010	AG	
RRIM2011	AG	
RRIM2012	AG	
RRIM2015	AG	
RRIM2017	AG	
RRIM2019	AG	
RRIM2020	AG	
RRIM2024	AG	
RRIM2025	AG	
RRIM3001	AG	
BD10	Undetermined	
PB260	Undetermined	
RRIM908	Undetermined	
RRIM2014	Undetermined	
RRIM2021	Undetermined	

REFERENCES

ABDEL-GHANY, S. E., HAMILTON, M., JACOBI, J. L., NGAM, P., DEVITT, N., SCHILKEY, F., BEN-HUR, A. & REDDY, A. S. N. 2016. A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications,* 7**,** 11706.

ABDELRAHMAN, M., JOGAIAH, S., BURRITT, D. J. & TRAN, L.-S. P. 2018. Legume genetic resources and transcriptome dynamics under abiotic stress conditions. *Plant, Cell & Environment,* 41**,** 1972-1983.

ABNIZOVA, I., TE BOEKHORST, R. & ORLOV, Y. L. 2017. Computationan errors and biases in short read next generation sequencing. *Journal of Proteomics & Bioinformatics***,** 1-17.

ABRAHAM, P. D. & HASHIM, I. 1983. Exploitation procedures for modern *Hevea* cultivars. *Proceedings of RRIM Planters' Conference.* Kuala Lumpur: Rubber Research Institute of Malaysia.

ABRAHAM, P. D., P'NG, T. C., LEE, C. K., SIVAKUMARAN, S., MANIKAM, B. & YEOH, C. P. Ethrel stimulation. Proceedings of the International Rubber Conference, 1975 Kuala Lumpur. Rubber Research Institute of Malaysia, 347-383.

AKHTAR, T. A., SUROWIECKI, P., SIEKIERSKA, H., KANIA, M., VAN GELDER, K., REA, K. A., VIRTA, L. K. A., VATTA, M., GAWARECKA, K., WOJCIK, J., DANIKIEWICZ, W., BUSZEWICZ, D., SWIEZEWSKA, E. & SURMACZ, L. 2017. Polyprenols Are Synthesized by a Plastidial cis-Prenyltransferase and Influence Photosynthetic Performance. *The Plant Cell,* 29**,** 1709-1725.

AKPO, E., CRANE, T. A., STOMPH, T.-J., TOSSOU, R. C., KOSSOU, D. K., VISSOH, P. V. & STRUIK, P. C. 2014. Social institutional dynamics of seed system reliability: the case of oil palm in Benin. *International Journal of Agricultural Sustainability,* 12**,** 214-232.

ALAGOZ, Y., NAYAK, P., DHAMI, N. & CAZZONELLI, C. I. 2018. *cis*-Carotene biosynthesis, evolution and regulation in plants: The emergence of novel signaling metabolites. *Archives of Biochemistry and Biophysics,* 654**,** 172-184.

ALLWOOD, J. W. May 2016 2016. *RE: Possible explanation of HILIC peaks with long tail (GGPP, GPP and FPP).*

ALPERT, A. J. 1990. Hydrophilic-interaction chromatography for the separation of peptides, nucleis acids and other polar compounds. *Journal of Chromatography,* 499**,** 177-196.

AMERIK, A. Y., MARTIROSYAN, Y. T. & GACHOK, I. V. 2018. Regulation of Natural Rubber Biosynthesis by Proteins Associated with Rubber Particles. *Russian Journal of Bioorganic Chemistry,* 44**,** 140-149.

AMPOMAH-DWAMENA, C., DRIEDONKS, N., LEWIS, D., SHUMSKAYA, M., CHEN, X., WURTZEL, E. T., ESPLEY, R. V. & ALLAN, A. C. 2015. The Phytoene synthase gene family of apple (Malus x domestica) and its role in controlling fruit carotenoid content. *BMC Plant Biology,* 15**,** 185.

ANASTASIO, A. E., PLATT, A., HORTON, M., GROTEWOLD, E., SCHOLL, R., BOREVITZ, J. O., NORDBORG, M. & BERGELSON, J. 2011. Source verification of mis-identified *Arabidopsis thaliana* accessions. *The Plant Journal,* 67**,** 554-566.

ANDREWS, S. C. 2015. *FastQC v0.11.3* [Online]. Cambridge, UK: Babraham Bioinformatics. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed].

APPELS, R., EVERSOLE, K., FEUILLET, C., KELLER, B., ROGERS, J., STEIN, N., POZNIAK, C. J., STEIN, N., CHOULET, F., DISTELFELD, A., EVERSOLE, K., POLAND, J., ROGERS, J., RONEN, G., SHARPE, A. G., POZNIAK, C., RONEN, G., STEIN, N., BARAD, O., BARUCH, K., CHOULET, F., KEEBLE-GAGNÈRE, G., MASCHER, M., SHARPE, A. G., BEN-ZVI, G., JOSSELIN, A.-A., STEIN, N., MASCHER, M., HIMMELBACH, A., CHOULET, F., KEEBLE-GAGNÈRE, G., MASCHER, M., ROGERS, J., BALFOURIER, F., GUTIERREZ-GONZALEZ, J., HAYDEN, M., JOSSELIN, A.-A., KOH, C., MUEHLBAUER, G., PASAM, R. K., PAUX, E., POZNIAK, C. J., RIGAULT, P., SHARPE, A. G., TIBBITS, J., TIWARI, V., CHOULET, F., KEEBLE-GAGNÈRE, G.,

MASCHER, M., JOSSELIN, A.-A., ROGERS, J., SPANNAGL, M., CHOULET, F., LANG, D., GUNDLACH, H., HABERER, G., KEEBLE-GAGNÈRE, G., MAYER, K. F. X., ORMANBEKOVA, D., PAUX, E., PRADE, V., ŠIMKOVÁ, H., WICKER, T., CHOULET, F., SPANNAGL, M., SWARBRECK, D., RIMBERT, H., FELDER, M., GUILHOT, N., GUNDLACH, H., HABERER, G., KAITHAKOTTIL, G., KEILWAGEN, J., LANG, D., LEROY, P., LUX, T., MAYER, K. F. X., TWARDZIOK, S., VENTURINI, L., APPELS, R., RIMBERT, H., CHOULET, F., JUHÁSZ, A., KEEBLE-GAGNÈRE, G., CHOULET, F., SPANNAGL, M., LANG, D., ABROUK, M., HABERER, G., KEEBLE-GAGNÈRE, G., MAYER, K. F. X., WICKER, T., CHOULET, F., WICKER, T., GUNDLACH, H., LANG, D., SPANNAGL, M., LANG, D., SPANNAGL, M., APPELS, R., et al. 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science,* 361.

APPULAGE, D. K. & SCHUG, K. A. 2017. Silica hydride based phases for small molecule separations using automated liquid chromatography–mass spectrometry method development. *Journal of Chromatography A,* 1507**,** 115-123.

ARCHER, B. L. & AUDLEY, B. G. 1987. New aspects of rubber biosynthesis. *Botanical Journal of the Linnean Society,* 94**,** 181-196.

ARCHER, B. L., AUDLEY, B. G., COCKBAIN, E. G. & MCSWEENEY, G. P. 1963. The biosynthesis of rubber. Incorporation of mevalonate and isopentenyl pyrophosphate into rubber by *Hevea brasiliensis*-latex fractions. *Biochemical Journal,* 89**,** 565-574.

ARCHER, B. L., AUDLEY, B. G., MCSWEENEY, G. P. & TAN, C. H. 1969. Studies on composition of latex serum and 'bottom fraction' particles. *Journal of Rubber Research Institute of Malaya,* 21**,** 560-569.

ARCHER, B. L., AYREY, G., COCKBAIN, E. G. & MCSWEENEY, G. P. 1961. Incorporation of [I-14C]-Isopentenyl Pyrophosphate into Polyisoprene. *Nature,* 189**,** 663.

ARCHER, B. L. & MCMULLEN, A. I. Some recent studies of the non-rubber constituents of of natural rubber latex. Proceedings of Natural Rubber Research Conference, 1960 Kuala Lumpur. The Rubber Research Institute of Malaya, 787-795.

ASAWATRERATANAKUL, K., ZHANG, Y. W., WITITSUWANNAKUL, D., WITITSUWANNAKUL, R., TAKAHASHI, S., RATTANAPITTAYAPORN, A. & KOYAMA, T. 2003. Molecular cloning, expression and characterisation of cDNA encoding *cis*-prenyltransferases from *Hevea brasiliensis*. *European Journal of Biochemistry,* 270**,** 4671-4680.

ASIF, M. J., ABDULLAH, M. Z., MUHAMMAD, N. & RATNAM, W. 2017. Detecting mislabeling and identifying unique progeny in *Acacia* mapping population using SNP markers. *Journal of Forestry Research,* 28**,** 1119-1127.

ATTANYAKA, D. P. S. T. G., KEKWICK, R. G. O. & FRANKLIN, F. C. H. 1991. Molecular cloning and nucleotide sequencing of the rubber elongation factor gene from *Hevea brasiliensis*. *Plant Molecular Biology,* 16**,** 1079-1081.

AU, K. F., SEBASTIANO, V., AFSHAR, P. T., DURRUTHY, J. D., LEE, L., WILLIAMS, B. A., VAN BAKEL, H., SCHADT, E. E., REIJO-PERA, R. A., UNDERWOOD, J. G. & WONG, W. H. 2013. Characterization of the human ESC transcriptome by hybrid sequencing. *Proceedings of the National Academy of Sciences,* 110**,** E4821-E4830.

BACKHAUS, R. A. 1985. Rubber formation in plants - a mini-review. *Israel Journal of Botany,* 34**,** 283-293.

BAHRI, A. R. S. & HAMZAH, S. 1996. Immunocytochemical localisation of rubber membrane protein in *Hevea* latex. *Journal of natural Rubber Research,* 11**,** 88-95.

BAJAD, S. U., LU, W., KIMBALL, E. H., YUAN, J., PETERSON, C. & RABINOWITZ, J. D. 2006. Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. *Journal of Chromatography A,* 1125**,** 76-88.

BALCKE, G., BENNEWITZ, S., BERGAU, N., ATHMER, B., HENNING, A., MAJOVSKY, P., JIMÉNEZ-GÓMEZ, J. M., HOEHENWARTER, W. & TISSIER, A. F. 2017. Multiomics of tomato glandular trichomes reveals distinct features of central carbon metabolism supporting high productivity of specialized metabolites. *The Plant Cell*.

BALCKE, G. U., BENNEWITZ, S., ZABEL, S. & TISSIER, A. 2014. Isoprenoid and metabolite profiling of plant trichomes. *In:* RODRÍGUEZ-CONCEPCIÓN, M. (ed.) *Plant Isoprenoids.* Springer New York.

BANDURSKI, R. S. & TEAS, H. J. 1957. Rubber biosynthesis in latex of *Hevea brasiliensis*. *Plant Physiology,* 32**,** 643-648.

BAUER, G., GORB, S. N., KLEIN, M.-C., NELLESEN, A., VON TAPAVICZA, M. & SPECK, T. 2014. Comparative study on plant latex particles and latex coagulation in *Ficus benjamina, Campanula glomerata* and three *Euphorbia* species. *PLOS ONE,* 9**,** e113336.

BEALING, F. J. 1969a. Carbohydrate metabolism in *Hevea* latex - availability and utilisation of substrates. *Journal of Rubber Research Institute of Malaya,* 21**,** 445-455.

BEALING, F. J. 1969b. Quantitative aspects of latex metabolism: possible involvement of precursors other that sucrose in the biosynthesis of *Hevea* rubber. *Proceedings of the International Rubber Conference.* Kuala Lumpur: The Rubber Research Institute of Malaysia.

BERTHELOT, K., LECOMTE, S., ESTEVEZ, Y., COULARY-SALIN, B., BENTALEB, A., CULLIN, C., DEFFIEUX, A. & PERUCH, F. 2012. Rubber elongation factor (REF), a major allergen component in *Hevea* brasiliensis latex has amyloid properties. *PLoS ONE,* 7**,** e48065. doi:10.1371/journal.pone.0048065.

BERTHELOT, K., LECOMTE, S., ESTEVEZ, Y. & PERUCH, F. 2014a. *Hevea* brasiliensis REF (Hev b 1) and SRPP (Hev b 3): An overview on rubber particles proteins. *Biochimie,* 106**,** 1-9.

BERTHELOT, K., LECOMTE, S., ESTEVEZ, Y., ZHENDRE, V., HENRY, S., THÉVENOT, J., DUFOURC, E. J., ALVES, I. D. & PERUCH, F. 2014b. Rubber particle proteins, HbREF and HbSRPP, show different interactions with model membranes. *Biochimica et Biophysica Acta (BBA) - Biomembranes,* 1838**,** 287-299.

BESSE, P., SEGUIN, M., LEBRUN, P., CHEVALLIER, M. H., NICOLAS, D. & LANAUD, C. 1994. Genetic diversity among wild and cultivated populations of *Hevea brasiliensis* assessed by nuclear RFLP analysis. *Theoretical and Applied Genetics,* 88**,** 199-207.

BEVAN, M. W., UAUY, C., WULFF, B. B. H., ZHOU, J., KRASILEVA, K. & CLARK, M. D. 2017. Genomic innovation for crop improvement. *Nature,* 543**,** 346.

BINO, R. J., HALL, R. D., FIEHN, O., KOPKA, J., SAITO, K., DRAPER, J., NIKOLAU, B. J., MENDES, P., ROESSNER-TUNALI, U., BEALE, M. H., TRETHEWEY, R. N., LANGE, B. M., WURTELE, E. S. & SUMNER, L. W. 2004. Potential of metabolomics as a functional genomics tool. *Trends in Plant Science,* 9**,** 418-425.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.

BONFILS, F., EHABE, E. E., AYMARD, C., VAYSSE, L. & SAINTE-BEUVE, J. 2007. Enhanced solvent extraction of polar lipids associated with rubber particles from *Hevea brasiliensis*. *Phytochemical Analysis,* 18**,** 103-108.

BOON, C. S., MCCLEMENTS, D. J., WEISS, J. & DECKER, E. A. 2010. Factors influencing the chemical stability of carotenoids in foods. *Critical Reviews in Food Science and Nutrition,* 50**,** 515-532.

BOONYANG, H. & SAKDAPIPANICH, J. 2014. The method to produce light-color natural rubber. *Advanced Materials Research,* 844**,** 85-88.

BOTELLA-PAVÍA, P., BESUMBES, Ó., PHILLIPS, M. A., CARRETERO-PAULET, L., BORONAT, A. & RODRÍGUEZ-CONCEPCIÓN, M. 2004. Regulation of carotenoid biosynthesis in plants: evidence for a key role of hydroxymethylbutenyl diphosphate reductase in controlling the supply of plastidial isoprenoid precursors. *The Plant Journal,* 40**,** 188-199.

BREDESON, J. V., LYONS, J. B., PROCHNIK, S. E., WU, G. A., HA, C. M., EDSINGER-GONZALES, E., GRIMWOOD, J., SCHMUTZ, J., RABBI, I. Y., EGESI, C., NAULUVULA, P., LEBOT, V., NDUNGURU, J., MKAMILO, G., BART, R. S., SETTER, T. L., GLEADOW, R. M., KULAKOW, P., FERGUSON, M. E., ROUNSLEY, S. & ROKHSAR, D. S. 2016. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nature Biotechnology,* 34**,** 562.

BRITTON, G. 1995. Structure and properties of carotenoids in relation to function. *The FASEB Journal,* 9**,** 1551-1558.

BROWN, D., FEENEY, M., AHMADI, M., LONOCE, C., SAJARI, R., DI COLA, A. & FRIGERIO, L. 2017a. Subcellular localization and interactions among rubber particle proteins from *Hevea brasiliensis*. *Journal of Experimental Botany,* 68**,** 5045-5055.

BROWN, J. W. S., CALIXTO, C. P. G. & ZHANG, R. 2017b. High-quality reference transcript datasets hold the key to transcript-specific RNA-sequencing analysis in plants. *New Phytologist,* 213**,** 525-530.

BROWN, M., DUNN, W. B., DOBSON, P., PATEL, Y., WINDER, C. L., FRANCIS-MCINTYRE, S., BEGLEY, P., CARROLL, K., BROADHURST, D., TSENG, A., SWAINSTON, N., SPASIC, I., GOODACRE, R. & KELL, D. B. 2009. Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst,* 134**,** 1322-1332.

BURANOV, A. U. & ELMURADOV, B. J. 2010. Extraction and characterization of latex and natural rubber from rubber-bearing plants. *Journal of Agricultural and Food Chemistry,* 58**,** 734-743.

BURKILL, H. M. 1959. Large scale variety trials of *Hevea brasiliensis* Muell. Arg. on Malayan estates 1934-53. *Journal of Rubber Research Institute of Malaya,* 16.

BUSZEWSKI, B. & NOGA, S. 2012. Hydrophilic interaction liquid chromatography (HILIC)—a powerful separation technique. *Analytical and Bioanalytical Chemistry,* 402**,** 231-247.

CAMPBELL, R., DUCREUX, L. J. M., MORRIS, W. L., MORRIS, J. A., SUTTLE, J. C., RAMSAY, G., BRYAN, G. J., HEDLEY, P. E. & TAYLOR, M. A. 2010. The metabolic and developmental roles of carotenoid cleavage dioxygenase4 from potato. *Plant Physiology,* 154**,** 656-664.

CAMPBELL, R., PONT, S. D. A., MORRIS, J. A., MCKENZIE, G., SHARMA, S. K., HEDLEY, P. E., RAMSAY, G., BRYAN, G. J. & TAYLOR, M. A. 2014. Genome-wide QTL and bulked transcriptomic analysis reveals new candidate genes for the control of tuber carotenoid content in potato (*Solanum tuberosum* L.). *Theoretical and Applied Genetics,* 127**,** 1917-1933.

CAZZONELLI, C. I. & POGSON, B. J. 2010. Source to sink: regulation of carotenoid biosynthesis in plants. *Trends in Plant Science,* 15**,** 266-274.

CHA, S. & YEUNG, E. S. 2007. Colloidal graphite-assisted laser desorption/ionization mass spectrometry and msn of small molecules. 1. Imaging of cerebrosides directly from rat brain tissue. *Analytical Chemistry,* 79**,** 2373-2385.

CHAIN, P. S. G., GRAFHAM, D. V., FULTON, R. S., FITZGERALD, M. G., HOSTETLER, J., MUZNY, D., ALI, J., BIRREN, B., BRUCE, D. C., BUHAY, C., COLE, J. R., DING, Y., DUGAN, S., FIELD, D., GARRITY, G. M., GIBBS, R., GRAVES, T., HAN, C. S., HARRISON, S. H., HIGHLANDER, S., HUGENHOLTZ, P., KHOURI, H. M., KODIRA, C. D., KOLKER, E., KYRPIDES, N. C., LANG, D., LAPIDUS, A., MALFATTI, S. A., MARKOWITZ, V., METHA, T., NELSON, K. E., PARKHILL, J., PITLUCK, S., QIN, X., READ, T. D., SCHMUTZ, J., SOZHAMANNAN, S., STERK, P., STRAUSBERG, R. L., SUTTON, G., THOMSON, N. R., TIEDJE, J. M., WEINSTOCK, G., WOLLAM, A. & DETTER, J. C. 2009. Genome project standards in a new era of sequencing. *Science,* 326**,** 236-237.

CHAKRABARTY, R., QU, Y. & RO, D. K. 2014. Silencingthe lettuce homologs of small rubber particle protein does not influence natural rubber biosynthesis in lettuce (*Lactuca sativa*). *Phytochemistry,* 113**,** 121-129.

CHAPMAN, K. D., DYER, J. M. & MULLEN, R. T. 2012. Biogenesis and functions of lipid droplets in plants: Thematic review series: Lipid droplet synthesis and metabolism: From yeast to man. *Journal of Lipid Research,* 53**,** 215-226.

CHAPPELL, J. 1995. Biochemistry and molecular biology of the isoprenoid biosynthetic pathway in plants. *Annual Review of Plant Physiology and Plant Molecular Biology,* 46**,** 521-547.

CHARBIT, E., LEGAVRE, T., LARDET, L., BOURGEOIS, E., FERRIÈRE, N. & CARRON, M. P. 2004. Identification of differentially expressed cDNA sequences and histological

characteristics of *Hevea brasiliensis* calli in relation to their embryogenic and regenerative capacities. *Plant Cell Reports,* 22**,** 539-548.

CHEN, S. F. 1981. Determination of particle size and specific surface area of particle in natural rubber latex. *1st Federation of Asian Chemical Society Regional Chemistry Seminaru and IKM Annual Chemical Conference.* Kuala Lumpur.

CHIANG, C. C. K., BARKAKATY, B., PUSKAS, J. E., XIE, W., CORNISH, K., PERUCH, F. & DEFFIEUX, A. 2014. Unraveling the mystery of natural rubber biosynthesis. Part II: Composition and growth of in vitro natural rubber using high-resolution size exclusion chromatography. *Rubber Chemistry and Technology,* 87**,** 451-458.

CHIN, H. C. & SINGH, M. M. 1980. Determination of dry rubber content of *Hevea* field latex. *Planters' Bulletin.* Kuala Lumpur, Malaysia: Rubber Research Institute of Malaysia.

CHOW, K. S., AHMAD-KAMAL, G., HOH, C. C. & ZAINORLINA, M. Z. 2014. RNA sequencing read depth requirement for optimal transcriptome coverage in *Hevea brasiliensis*. *BMC Research Notes,* 7.

CHOW, K. S., MOHD-NOOR, M.-I., BAHARI, A., AHMAD-KAMAL, G., ALIAS, H., ZAINORLINA, M.-Z., HOH, C.-C. & WAN, K.-L. 2012. Metabolic routes affecting rubber biosynthesis in *Hevea* brasiliensis latex. *Journal of Experimental Botany,* 63**,** 1863-1871.

CHOW, K. S., WAN, K.-L., MAT-ISA, M.-N., BAHARI, A., TAN, S.-H., HARIKRISHNA, K. & YEANG, H. Y. 2007. Insights into rubber biosynthesis from transcriptome analysis of *Hevea brasiliensis* latex. *Journal of Experimental Botany,* 58**,** 2429-2440.

CHOW, K. S., YUSOF, F., MOHD-LIM, S. & ABDULLAH, L. 2006. An assessment of stimulation of *Hevea brasiliensis* rubber biosynthesis by eIF-5A protein-enriched bacterial lysates. *Journal of Rubber Research,* 9**,** 251-259.

CLAVIJO, B. J., VENTURINI, L., SCHUDOMA, C., ACCINELLI, G. G., KAITHAKOTTIL, G., WRIGHT, J., BORRILL, P., KETTLEBOROUGH, G., HEAVENS, D., CHAPMAN, H., LIPSCOMBE, J., BARKER, T., LU, F.-H., MCKENZIE, N., RAATS, D., RAMIREZ-GONZALEZ, R. H., COINCE, A., PEEL, N., PERCIVAL-ALWYN, L., DUNCAN, O., TRÖSCH, J., YU, G., BOLSER, D. M., NAMAATI, G., KERHORNOU, A., SPANNAGL, M., GUNDLACH, H., HABERER, G., DAVEY, R. P., FOSKER, C., PALMA, F. D., PHILLIPS, A., MILLAR, A. H., KERSEY, P. J., UAUY, C., KRASILEVA, K. V., SWARBRECK, D., BEVAN, M. W. & CLARK, M. D. 2017. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research*.

COCKBAIN, E. G. 1953. Determination of particle size in latex. *Rubber Chemistry and Technology,* 26**,** 481-486.

COCKBAIN, E. G. & PHILPOTT, M. W. 1963. *The chemistry and physics of rubber-like substances,* London, Maclaren and Sons Ltd.

COCKBAIN, E. G. & SOUTHORN, W. A. 1962. The structure and composition of *Hevea*. *Rev. Gen. Caoutch.***,** 1149-1156.

COLASUONNO, P., LOZITO, M. L., MARCOTULI, I., NIGRO, D., GIANCASPRO, A., MANGINI, G., DE VITA, P., MASTRANGELO, A. M., PECCHIONI, N., HOUSTON, K., SIMEONE, R., GADALETA, A. & BLANCO, A. 2017. The carotenoid biosynthetic and catabolic genes in wheat and their association with yellow pigments. *BMC Genomics,* 18**,** 122.

COLLINS-SILVA, J., NURAL, A. T., SKAGGS, A., SCOTT, D., HATHWAIK, U., WOOLSEY, R., SCHEGG, K., MCMAHAN, C., WHALEN, M., CORNISH, K. & SHINTANI, D. 2012. Altered levels of the *Taraxacum koksaghyz* (Russian dandelion) small rubber particle protein, *TkSRPP3*, result in qualitative and quantitative changes in rubber metabolism. *Phytochemistry,* 79**,** 46-56.

CONSON, A. R. O., TANIGUTI, C. H., AMADEU, R. R., ANDREOTTI, I. A. A., DE SOUZA, L. M., DOS SANTOS, L. H. B., ROSA, J. R. B. F., MANTELLO, C. C., DA SILVA, C. C., JOSÉ SCALOPPI JUNIOR, E., RIBEIRO, R. V., LE GUEN, V., GARCIA, A. A. F., GONÇALVES, P. D. S. & DE SOUZA, A. P. 2018. High-resolution genetic map and qtl analysis of growth-related

traits of *Hevea brasiliensis* cultivated under suboptimal temperature and humidity conditions. *Frontiers in Plant Science,* 9.

COOK, A. M. 1960. The short-term preservation of natural latex *Journal of Rubber Research Institute of Malaya,* 16**,** 65.

CORNISH, K. 2001. Similarities and differences in rubber biochemistry among plant species. *Phytochemistry,* 57**,** 1123-1134.

CORNISH, K. & BLAKESLEE, J. 2011. Rubber biosynthesis in plants.

CORNISH, K. & BRICHTA, J. L. 2002. Some rheological properties of latex from *Parthenium argentatum* Gray compared with latex from *Hevea brasiliensis* and *Ficus elastica*. *Journal of Polymers and the Environment,* 10**,** 13-18.

CORNISH, K. & SILER, D. J. 1995. Effect of Different Allylic Diphosphates on the Initiation of New Rubber Molecules and on Cis-1,4-polyisoprene Biosynthesis in Guayule (Parthenium argentatum Gray). *Journal of Plant Physiology,* 147**,** 301-305.

CORNISH, K., WOOD, D. F. & WINDLE, J. J. 1999. Rubber particles from four different species, examined by transmission electron microscopy and electron-paramagnetic-resonance spin labeling, are found to consist of a homogeneous rubber core enclosed by a contiguous, monolayer biomembrane. *Planta,* 210**,** 85-96.

CUNNINGHAM, F. X. & GANTT, E. 2001. One ring or two? Determination of ring number in carotenoids by lycopene ε-cyclases. *Proceedings of the National Academy of Sciences,* 98**,** 2905-2910.

D'AUZAC, J. & JACOB, J. L. 1989. The composition of latex from *Hevea* brasiliensis as a laticferous cytoplasm. *In:* D'AUZAC, J., JACOB, J. L. & CHRESTIN, H. (eds.) *Physiology of Rubber Tree Latex.* Boca Raton, Florida, US: CRC Press Inc.

DAI, L., KANG, G., NIE, Z., DUAN, C. & ZENG, R. 2013. In-depth proteome analysis of the rubber particle of *Hevea brasiliensis* (para rubber tree). *Plant Molecular Biology,* 82**,** 155-168.

DAI, L., NIE, Z., KANG, G., LI, Y. & ZENG, R. 2017. Identification and subcellular localization analysis of two rubber elongation factor isoforms on *Hevea brasiliensis* rubber particles. *Plant Physiology and Biochemistry,* 111**,** 97-106.

DENNIS, M. S., HENZEL, W. J., BELL, J., KOHR, W. & LIGHT, D. R. 1989. Amino acid sequence of rubber elongation factor protein associated with rubber particles in *Hevea* latex. *Journal of Biological Chemistry,* 264**,** 18618-18626.

DEVARAJ, V., NUR-FADHILAH, I., NOR-HIDAYATI, K. & ZAIROSSANI, M. N. 2013. SMR malodor and environmental issues. *MRB Rubber Technology Developments.* Kuala Lumpur: Malaysian Rubber Bpard.

DHANDAPANI, R., SINGH, V. P., ARORA, A., BHATTACHARYA, R. C. & RAJENDRAN, A. 2017. Differential accumulation of β-carotene and tissue specific expression of phytoene synthase (*MaPsy*) gene in banana (*Musa* sp) cultivars. *Journal of Food Science and Technology,* 54**,** 4416-4426.

DHAR, M. K., SHARMA, R., KOUL, A. & KAUL, S. 2015. Development of fruit color in *Solanaceae*: a story of two biosynthetic pathways. *Briefings in Functional Genomics,* 14**,** 199-212.

DICKENSON, P. B. 1969. Electron microscopical studies of the latex vessel system of *Hevea brasiliensis*. *Journal of Rubber Research Institute of Malaya,* 21**,** 543.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29**,** 15-21.

DOMINGUES, M. R. M., REIS, A. & DOMINGUES, P. 2008. Mass spectrometry analysis of oxidized phospholipids. *Chemistry and Physics of Lipids,* 156**,** 1-12.

DROZDETSKIY, A., COLE, C., PROCTER, J. & BARTON, G. J. 2015. JPred4: a protein secondary structure prediction server. *Nucleic Acids Research,* 43**,** W389-W394.

DUCREUX, L. J. M., MORRIS, W. L., HEDLEY, P. E., SHEPHERD, T., DAVIES, H. V., MILLAM, S. & TAYLOR, M. A. 2005. Metabolic engineering of high carotenoid potato tubers containing enhanced levels of β-carotene and lutein. *Journal of Experimental Botany,* 56**,** 81-89.

DULNGALI, S. & ONG, C. O. 1980. Production of light-coloured rubbers - SMR EQ, L and CV. *RRIM Training Manual on Natural Rubber Processing.* Kuala Lumpur, Malaysia: Rubber Research Institute of Malaysia.

DUNPHY, P. J., WHITTLE, K. J. & MORTON, R. A. 1965. Identification and estimation of tocotrienols in *Hevea* latex. *Nature,* 207**,** 521-522.

DUSOTOIT-COUCAUD, A., BRUNEL, N., KONGSAWADWORAKUL, P., VIBOONJUN, U., LACOINTE, A., JULIEN, J.-L., CHRESTIN, H. & SAKR, S. 2009. Sucrose importation into laticifers of *Hevea brasiliensis*, in relation to ethylene stimulation of latex production. *Annals of Botany,* 104**,** 635-647.

EATON, B. J. & FULLERTON, G. 1929. The identification and quantitative estimation of the yellow pigment of raw rubber. *Journal of Rubber Research Institute of Malaya* 1**,** 135-148.

EHABE, E. E. & BONFILS, F. 2011. Novel insight into the gel phase of *Hevea* natural rubber. *Journal of Rubber Research,* 14**,** 1-10.

ELLISON, S., SENALIK, D., BOSTAN, H., IORIZZO, M. & SIMON, P. 2017. Fine mapping, transcriptome analysis, and marker development for Y_2, the gene that conditions beta-carotene accumulation in carrot (*Daucus carota* L.). *G3: Genes|Genomes|Genetics*.

ENFISSI, E. M. A., NOGUEIRA, M., BRAMLEY, P. M. & FRASER, P. D. 2017. The regulation of carotenoid formation in tomato fruit. *The Plant Journal,* 89**,** 774-788.

ENG, A. H., OTHMAN, H., HASMA, H., RAMLI, O., MASAHULING, B., MUNIANDY, V. & KAWAHARA, S. 2001. Some properties of natural rubber from latex-timber clones. *Jounral of Rubber Research,* 4**,** 164-176.

FENG, S. P., LI, W. G., HUANG, H. S., WANG, J. Y. & WU, Y. T. 2009. Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (*Hevea brasiliensis*). *Molecular Breeding,* 23**,** 85-97.

FINKELSTEIN, R. R., GAMPALA, S. S. L. & ROCK, C. D. 2002. Abscisic acid signaling in seeds and seedlings. *The Plant Cell,* 14**,** S15-S45.

FLAGEL, L. E. & WENDEL, J. F. 2009. Gene duplication and evolutionary novelty in plants. *New Phytologist,* 183**,** 557-564.

FOLCH, J., LEES, M. & STANLEY, G. 1956. A simple method for the isolation and purification of total lipids from animal tissues. *Journal of Biological Chemistry***,** 497-509.

FRASER, P. D. & BRAMLEY, P. M. 2004. The biosynthesis and nutritional uses of carotenoids. *Progress in Lipid Research,* 43**,** 228-265.

FRASER, P. D., PINTO, M. E. S., HOLLOWAY, D. E. & BRAMLEY, P. M. 2000. Application of high-performance liquid chromatography with photodiode array detection to the metabolic profiling of plant isoprenoids. *The Plant Journal,* 24**,** 551-558.

FRASER, P. D., PINTO, M. E. S., HOLLOWAY, D. E. & BRAMLEY, P. M. 2008. Application of high-performance liquid chromatography with photodiode array detection to the metabolic profiling of plant isoprenoids. *The Plant Journal,* 24**,** 551-558.

GADY, A. L. F., VRIEZEN, W. H., VAN DE WAL, M. H. B. J., HUANG, P., BOVY, A. G., VISSER, R. G. F. & BACHEM, C. W. B. 2012. Induced point mutations in the phytoene synthase 1 gene cause differences in carotenoid content during tomato fruit ripening. *Molecular Breeding,* 29**,** 801-812.

GARRISON, E. & MARTH, G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv*.

GIDDA, S. K., PARK, S., PYC, M., YURCHENKO, O., CAI, Y., WU, P., ANDREWS, D. W., CHAPMAN, K. D., DYER, J. M. & MULLEN, R. T. 2016. Lipid Droplet-Associated Proteins (LDAPs) Are Required for the Dynamic Regulation of Neutral Lipid Compartmentation in Plant Cells. *Plant Physiology,* 170**,** 2052-2071.

GIKA, H. G., THEODORIDIS, G. A., VRHOVSEK, U. & MATTIVI, F. 2012. Quantitative profiling of polar primary metabolites using hydrophilic interaction ultrahigh performance liquid chromatography–tandem mass spectrometry. *Journal of Chromatography A,* 1259**,** 121-127.

GIRAUDEAU, P. 2017. Challenges and perspectives in quantitative NMR. *Magnetic Resonance in Chemistry,* 55**,** 61-69.

GIULIANO, G., TAVAZZA, R., DIRETTO, G., BEYER, P. & TAYLOR, M. A. 2008. Metabolic engineering of carotenoid biosynthesis in plants. *Trends in Biotechnology,* 26**,** 139-145.

GOMEZ, J. B. 1990. Lutoids of *Hevea* latex: morphological considerations. *Journal of natural Rubber Research,* 5**,** 231-240.

GOMEZ, J. B. & MOIR, G. F. J. 1979. Ultracytology of latex vessels in *Hevea brasiliensis. MRRDB Monograph No. 4.* Kuala Lumpur, Malaysia: Malaysian Rubber Research and Development Board.

GOMEZ, J. B. & SAMSIDAR, H. 1989. *Frey-Wyssling* complex in *Hevea* latex - uniqueness of the organelle. *Journal of natural Rubber Research,* 4**,** 75-85.

GUO, D., ZHOU, Y., LI, H.-L., ZHU, J.-H., WANG, Y., CHEN, X.-T. & PENG, S.-Q. 2017. Identification and characterization of the abscisic acid (ABA) receptor gene family and its expression in response to hormones in the rubber tree. *Scientific Reports,* 7**,** 45157.

GUO, Y. & GAIKI, S. 2005. Retention behavior of small polar compounds on polar stationary phases in hydrophilic interaction chromatography. *Journal of Chromatography A,* 1074**,** 71-80.

GUTTMAN, M., GARBER, M., LEVIN, J. Z., DONAGHEY, J., ROBINSON, J., ADICONIS, X., FAN, L., KOZIOL, M. J., GNIRKE, A., NUSBAUM, C., RINN, J. L., LANDER, E. S. & REGEV, A. 2010. Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology,* 28**,** 503.

HAAS, B. J. 2016. *RE: Default setting of Trinity assembler*

HACKL, T., HEDRICH, R., SCHULTZ, J. & FÖRSTER, F. 2014. proovread : large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics,* 30**,** 3004-3011.

HAGEL, J. M., YEUNG, E. C. & FACCHINI, P. J. 2008. Got milk? The secret life of laticifers. *Trends in Plant Science,* 13**,** 631-639.

HAN, K.-H., SHIN, D. H., YANG, J., KIM, I. J., OH, S. K. & CHOW, K. S. 2000. Genes expressed in the latex of *Hevea* brasiliensis. *Tree Physiology,* 20**,** 503-510.

HAN, X.-J., WANG, Y.-D., CHEN, Y.-C., LIN, L.-Y. & WU, Q.-K. 2013. Transcriptome sequencing and expression analysis of terpenoid biosynthesis genes in *Litsea cubeba*. *PLOS ONE,* 8**,** e76890.

HAN, Y., GAO, S., MUEGGE, K., ZHANG, W. & ZHOU, B. 2015. Advanced applications of rna sequencing and challenges. *Bioinformatics and Biology Insights,* 9**,** 29-46.

HAO, B. Z. & WU, J. L. 2000. Laticifer differentiation in *Hevea brasiliensis*: Induction by exogenous jasmonic acid and linolenic acid. *Annals of Botany,* 85**,** 37-43.

HARJU, A. & MUONA, O. 1989. Background pollination in *Pinus sylvestris* seed orchards. *Scandinavian Journal of Forest Research,* 4**,** 513-520.

HASMA, H. 1991. Lipids associated with rubber particles and their possible role in mechanical stability of latex concentrates. *Journal of natural Rubber Research,* 6**,** 105-114.

HASMA, H. & ALIAS, O. 1990. Role of some non-rubber constituents on thermal oxidative ageing of natural rubber. *Journal of Natural Rubber Research,* 5**,** 1-8.

HASMA, H. & SUBRAMANIAM, A. 1986. Composition of lipids in latex of *Hevea brasiliensis* clone RRIM 501. *Journal of natural Rubber Research,* 1**,** 30-40.

HASUNUMA, T., HARADA, K., MIYAZAWA, S.-I., KONDO, A., FUKUSAKI, E. & MIYAKE, C. 2010. Metabolic turnover analysis by a combination of in vivo13C-labelling from 13CO2 and metabolic profiling with CE-MS/MS reveals rate-limiting steps of the C3 photosynthetic pathway in Nicotiana tabacum leaves. *Journal of Experimental Botany,* 61**,** 1041-1051.

HAWKINS, W. L. 1984. Polymer Degradation. *In:* HAWKINS, W. L. (ed.) *Polymer Degradation and Stabilization.* Berlin, Heidelberg: Springer Berlin Heidelberg.

HEATHER, J. M. & CHAIN, B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics,* 107**,** 1-8.

HEMMERLIN, A., HARWOOD, J. L. & BACH, T. J. 2012. A raison d'être for two distinct pathways in the early steps of plant isoprenoid biosynthesis? *Progress in Lipid Research,* 51**,** 95-148.

HENNEMAN, L., VAN CRUCHTEN, A. G., DENIS, S. W., AMOLINS, M. W., PLACZEK, A. T., GIBBS, R. A., KULIK, W. & WATERHAM, H. R. 2008. Detection of nonsterol isoprenoids by HPLC-MS/MS. *Analytical Biochemistry,* 383**,** 18-24.

HEPPER, C. M. & AUDLEY, B. G. 1969. The biosynthesis of rubber from β-hydroxy-β-methylglutaryl-coenzyme A in *Hevea brasiliensis* latex. *Biochemistry Journal,* 144**,** 370-386.

HILLEBRAND, A., POST, J. J., WURBS, D., WAHLER, D., LENDERS, M., KRZYZANEK, V., PRÜFER, D. & GRONOVER, C. S. 2012. Down-regulation of small rubber particle protein expression affects integrity of rubber particles and rubber content in *Taraxacum brevicorniculatum*. *PLoS ONE,* 7**,** e41874. doi:10.1371/journal.pone.0041874.

HIRSCHBERG, J. 2001. Carotenoid biosynthesis in flowering plants. *Current Opinion in Plant Biology,* 4**,** 210-218.

HO, C. C. 1989. Changes in electrokinetic properties of natural rubber latex after surface chemical modifications. *Colloid and Polymer Science,* 267**,** 643-647.

HO, C. C., SUBRAMANIAM, A. & YONG, W. M. 1975a. Lipids associated with the particles in *Hevea* latex. *Proceedings of the International Rubber Conference.* Kuala Lumpur: The Rubber Research Institute of Malaysia.

HO, C. C., SUBRAMANIAM, A. & YONG, W. M. 1975b. Lipids associated with the particles in *Hevea* latex. *Proceedings of the International Rubber Conference.* Kuala Lumpur: The Rubber Research Institute of Malaysia.

HORN, P. J., JAMES, C. N., GIDDA, S., KILARU, A., DYER, J. M., MULLEN, R. T., OHLROGGE, J. B. & CHAPMAN, K. 2013. Identification of a new class of lipid droplet-associated proteins in plants. *Plant Physiology*.

HOWITT, C. A. & POGSON, B. J. 2006. Carotenoid accumulation and function in seeds and non-green tissues. *Plant, Cell & Environment,* 29**,** 435-445.

HRDLICKOVA, R., TOLOUE, M. & TIAN, B. 2017. RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA,* 8**,** e1364.

HSIEH, Y. 2008. Potential of HILIC-MS in quantitative bioanalysis of drugs and drug metabolites. *Journal of Separation Science,* 31**,** 1481-1491.

HUH, J. H., KANG, B. C., NAHM, S. H., KIM, S., HA, K. S., LEE, M. H. & KIM, B. D. 2001. A candidate gene approach identified phytoene synthase as the locus for mature fruit color in red pepper (*Capsicum* spp.). *Theoretical and Applied Genetics,* 102**,** 524-530.

HULSE-KEMP, A. M., MAHESHWARI, S., STOFFEL, K., HILL, T. A., JAFFE, D., WILLIAMS, S. R., WEISENFELD, N., RAMAKRISHNAN, S., KUMAR, V., SHAH, P., SCHATZ, M. C., CHURCH, D. M. & VAN DEYNZE, A. 2018. Reference quality assembly of the 3.5-Gb genome of *Capsicum annuum* from a single linked-read library. *Horticulture Research,* 5**,** 4.

HUNTER, J. R. 1994. Reconsidering the functions of latex. *Trees,* 9**,** 1-5.

HURTADO-PÁEZ, U. A., GARCÍA ROMERO, I. A., RESTREPO RESTREPO, S., ARISTIZÁBAL GUTIÉRREZ, F. A. & MONTOYA CASTAÑO, D. 2015. Assembly and analysis of differential transcriptome responses of *Hevea brasiliensis* on interaction with *Microcyclus ulei*. *PLOS ONE,* 10**,** e0134837.

HUSON, D. H. & SCORNAVACCA, C. 2012. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Systematic Biology,* 61**,** 1061-1067.

INBARAJ, B. S., LU, H., HUNG, C. F., WU, W. B., LIN, C. L. & CHEN, B. H. 2008. Determination of carotenoids and their esters in fruits of *Lycium barbarum* Linnaeus by HPLC–DAD–APCI–MS. *Journal of Pharmaceutical and Biomedical Analysis,* 47**,** 812-818.

ITO, J., HERTER, T., BAIDOO, E. E. K., LAO, J., VEGA-SÁNCHEZ, M. E., MICHELLE SMITH-MORITZ, A., ADAMS, P. D., KEASLING, J. D., USADEL, B., PETZOLD, C. J. & HEAZLEWOOD, J. L. 2014. Analysis of plant nucleotide sugars by hydrophilic interaction liquid chromatography and tandem mass spectrometry. *Analytical Biochemistry,* 448**,** 14-22.

JACOB, J. L., D'AUZAC, J. & PREVÔT, J. C. 1993. The composition of natural latex from *Hevea brasiliensis*. *Clinical Reviews in Allergy and Immunology,* 11**,** 325-337.

JAYASHREE, R., NAZEEM, P. A., REKHA, K., SREELATHA, S., THULASEEDHARAN, A., KRISHNAKUMAR, R., KALA, R. G., VINEETHA, M., LEDA, P., JINU, U. & VENKATACHALAM, P. 2018. Over-expression of 3-hydroxy-3- methylglutaryl-coenzyme A reductase 1 (*hmgr1*) gene under super-promoter for enhanced latex biosynthesis in rubber tree (*Hevea brasiliensis* Muell. Arg.). *Plant Physiology and Biochemistry,* 127**,** 414-424.

JI, W., BENEDICT, C. R. & FOSTER, M. A. 1993. Seasonal variations in rubber biosynthesis, 3-hydroxy-3-methylglutaryl-coenzyme a reductase, and rubber transferase activities in *Parthenium argentatum* in the Chihuahuan desert. *Plant Physiology,* 103**,** 535-542.

JOHN, C. K. 1976. Studies on some anticoagulants and preservatives of *Hevea* latex. *Journal of Rubber Research Institute of Malaysia,* 24**,** 137-144.

KACHANOVSKY, D. E., FILLER, S., ISAACSON, T. & HIRSCHBERG, J. 2012. Epistasis in tomato color mutations involves regulation of *phytoene synthase 1* expression by *cis*-carotenoids. *Proceedings of the National Academy of Sciences,* 109**,** 19021-19026.

KAJITANI, R., TOSHIMOTO, K., NOGUCHI, H., TOYODA, A., OGURA, Y., OKUNO, M., YABANA, M., HARADA, M., NAGAYASU, E., MARUYAMA, H., KOHARA, Y., FUJIYAMA, A., HAYASHI, T. & ITOH, T. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research,* 24**,** 1384-1395.

KAJIURA, H., SUZUKI, N., MOURI, H., WATANABE, N. & NAKAZAWA, Y. 2018. Elucidation of rubber biosynthesis and accumulation in the rubber producing shrub, guayule (*Parthenium argentatum* Gray). *Planta,* 247**,** 513-526.

KAO, T. H., LOH, C. H., INBARAJ, B. S. & CHEN, B. H. 2012. Determination of carotenoids in *Taraxacum formosanum* by HPLC–DAD–APCI-MS and preparation by column chromatography. *Journal of Pharmaceutical and Biomedical Analysis,* 66**,** 144-153.

KAWAUCHI, H. & GOTO, S. 1999. Monitoring clonal management in nematode-resistant Japanese black pine seed orchard in Kagoshima Prefecture. *JOURNAL OF THE JAPANESE FORESTRY SOCIETY,* 81**,** 338-340.

KEIL, M. & GRIFFIN, A. R. 1994. Use of random amplified polymorphic DNA (RAPD) markers in the discrimination and verification of genotypes in Eucalyptus. *Theoretical and Applied Genetics,* 89**,** 442-450.

KEKWICK, R., BHAMBRI, S., CHABANE, M. H., AUTEGARDEN, J.-E., LEVY, D. A. & LEYNADIER, F. 1996. The allergenic properties of fresh and preserved *Hevea brasiliensis* latex protein preparations. *Clinical & Experimental Immunology,* 104**,** 337-342.

KIM, E. Y., PARK, K. Y., SEO, Y. S. & KIM, W. T. 2016. *Arabidopsis* small rubber particle protein homolog srps play dual roles as positive factors for tissue growth and development and in drought stress responses. *Plant Physiology,* 170**,** 2494-2510.

KIM, J. & DELLAPENNA, D. 2006. Defining the primary route for lutein synthesis in plants: The role of *Arabidopsis* carotenoid β-ring hydroxylase CYP97A3. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 3474-3479.

KIMURA, M., RODRIGUEZ-AMAYA, D. B. & GODOY, H. T. 1990. Assessment of the saponification step in the quantitative determination of carotenoids and provitamins A. *Food Chemistry,* 35**,** 187-195.

KITAOKA, M., NAGAKI, H., KINOSHITA, T., KURABAYASHI, M., KOYAMA, T. & OGURA, K. 1990. Negative ion fast atom bombardment-tandem mass spectrometry for structural analysis of isoprenoid diphosphates. *Analytical Biochemistry,* 185**,** 182-186.

KO, J. H., CHOW, K.-S. & HAN, K. H. 2003. Transcriptome analysis reveals novel features of the molecular events occurring in the laticifers of *Hevea brasiliensis* (para rubber tree). *Plant Molecular Biology,* 53**,** 479-492.

KÖHLING, R., MEIER, R., SCHÖNENBERGER, B. & WOHLGEMUTH, R. 2015. Analysis of Isoprenoid Pathway Metabolites by LC-MS. *Biofiles* [Online], 2018. Available:

https://www.sigmaaldrich.com/technical-documents/articles/biofiles/isoprenoid-pathway-metabolites.html.

KOONIN, E. V. & WOLF, Y. I. 2010. Constraints and plasticity in genome and molecular-phenome evolution. *Nature Reviews. Genetics,* 11**,** 487-498.

KUO, R. I., TSENG, E., EORY, L., PATON, I. R., ARCHIBALD, A. L. & BURT, D. W. 2017. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics,* 18**,** 323.

KUSH, A., GOYVAERTS, E., CHYE, M. L. & CHUA, N. H. 1990. Laticifer-specific gene expression in *Hevea brasiliensis* (rubber tree). *Proc Natl Acad Sci U S A,* 87**,** 1787-90.

LACKER, T., STROHSCHEIN, S. & ALBERT, K. 1999. Separation and identification of various carotenoids by C30 reversed-phase high-performance liquid chromatography coupled to UV and atmospheric pressure chemical ionization mass spectrometric detection. *Journal of Chromatography A,* 854**,** 37-44.

LAIBACH, N., HILLEBRAND, A., TWYMAN, R. M., PRÜFER, D. & GRONOVER, C. S. 2015. Identification of a *Taraxacum brevicorniculatum* rubber elongation factor protein that is localised on rubber particles and promotes rubber biosynthesis. *The Plant Journal***,** doi:10.1111/tpj.12836.

LAIBACH, N., SCHMIDL, S., MÜLLER, B., BERGMANN, M., PRÜFER, D. & SCHULZE GRONOVER, C. 2018. Small rubber particle proteins from *Taraxacum brevicorniculatum* promote stress tolerance and influence the size and distribution of lipid droplets and artificial poly(cis-1,4-isoprene) bodies. *The Plant Journal,* 93**,** 1045-1061.

LANGE, B. M. 2015. The evolution of plant secretory structures and the emergence of terpenoid chemical diversity. *Annual Review of Plant Biology,* 66**,** 19.1-19.21.

LARSEN, E. & CHRISTENSEN, L. P. 2005. Simple saponification method for the quantitative determination of carotenoids in green vegetables. *Journal of Agricultural and Food Chemistry,* 53**,** 6598-6602.

LASHBROOKE, J. G., YOUNG, P. R., STANDER, C. & VIVIER, M. A. 2010. The development of a methof for the extraction of carotenoids and chlorophylls from grapevine leaves and berries for HPLC profiling. *Australian Journal of Grape and Wine Research,* 16**,** 349-360.

LAU, N.-S., MAKITA, Y., KAWASHIMA, M., TAYLOR, T. D., KONDO, S., OTHMAN, A. S., SHU-CHIEN, A. C. & MATSUI, M. 2016. The rubber tree genome shows expansion of gene family associated with rubber biosynthesis. *Scientific Reports,* 6**,** 28594.

LAW, C. W., ALHAMDOOSH, M., SU, S., SMYTH, G. K. & RITCHIE, M. E. 2016. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research,* 5**,** 1408.

LE GUEN, V., DOARÉ, F., WEBER, C. & SEGUIN, M. 2009. Genetic structure of Amazonian populations of *Hevea brasiliensis* is shaped by hydrographical network and isolation by distance. *Tree Genetics & Genomes,* 5**,** 673-683.

LEITCH, A. R., LIM, K. Y., LEITCH, I. J., O'NEILL, M., CHYE, M. & LOW, F. 1998. Molecular cytogenetic studies in rubber, *Hevea brasiliensis* Muell. Arg. (Euphorbiaceae). *Genome,* 41**,** 464-467.

LESPINASSE, D., GRIVET, L., TROISPOUX, V., RODIER-GOUD, M., PINARD, F. & SEGUIN, M. 2000. Identification of QTLs involved in the resistance to South American leaf blight (*Microcyclus ulei*) in the rubber tree. *Theoretical and Applied Genetics,* 100**,** 975-984.

LEWINSOHN, T. M. 1991. The geographical distribution of plant latex. *CHEMOECOLOGY,* 2**,** 64-68.

LI, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics***,** bty191-bty191.

LI, L., PAOLILLO, D. J., PARTHASARATHY, M. V., DIMUZIO, E. M. & GARVIN, D. F. 2001. A novel gene mutation that confers abnormal patterns of β-carotene accumulation in cauliflower (*Brassica oleracea* var. botrytis). *The Plant Journal,* 26**,** 59-67.

LI, L. & YUAN, H. 2013. Chromoplast biogenesis and carotenoid accumulation. *Archives of Biochemistry and Biophysics,* 539**,** 102-109.

LI, Z. & SHARKEY, T. D. 2013. Metabolic profiling of the methylerythritol phosphate pathway reveals the source of post-illumination isoprene burst from leaves. *Plant, Cell and Environment,* 36**,** 429-437.

LIAO, Y., SMYTH, G. K. & SHI, W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics,* 30**,** 923-930.

LICHTENTHALER, H. K. 1999. The 1-deoxy-d-xylulose-5-phosphate pathway of isoprenoid biosynthesis in plants. *Annual Review of Plant Physiology and Plant Molecular Biology,* 50**,** 47-65.

LIEBEREI, R. 2007. South American leaf blight of the rubber tree (*Hevea* spp.): New steps in plant domestication using physiological features and molecular markers. *Annals of Botany,* 100**,** 1125-1142.

LIENGPRAYOON, S., BONFILS, F., SAINTE-BEUVE, J., SRIROTH, K., DUBREUCQ, E. & VAYSSE, L. 2008. Development of a new procedure for lipid extraction from *Hevea brasiliensis* natural rubber. *European Journal of Lipid Science and Technology,* 110**,** 563-569.

LIENGPRAYOON, S., CHAIYUT, J., SRIROTH, K., BONFILS, F., SAINTE-BEUVE, J., DUBREUCQ, E. & VAYSSE, L. 2013. Lipids compositions of latex and sheet rubber from *Hevea brasiliensis* depend on clonal origin. *European Journal of Lipid Science and Technology,* 115**,** 1021-1031.

LIGHT, D. R. & DENNIS, M. S. 1989. Purification of a prenyltransferase that elongates cis-polyisoprene rubber from the latex of *Hevea brasiliensis*. *Journal of Biological Chemistry,* 264**,** 18589-18597.

LIN, T., XU, X., RUAN, J., LIU, S., WU, S., SHAO, X., WANG, X., GAN, L., QIN, B., YANG, Y., CHENG, Z., YANG, S., ZHANG, Z., XIONG, G., HUANG, S., YU, H. & LI, J. 2018. Genome analysis of *Taraxacum kok-saghyz* Rodin provides new insights into rubber biosynthesis. *National Science Review,* 5**,** 78-87.

LIN, X., KAUL, S., ROUNSLEY, S., SHEA, T. P., BENITO, M.-I., TOWN, C. D., FUJII, C. Y., MASON, T., BOWMAN, C. L., BARNSTEAD, M., FELDBLYUM, T. V., BUELL, C. R., KETCHUM, K. A., LEE, J., RONNING, C. M., KOO, H. L., MOFFAT, K. S., CRONIN, L. A., SHEN, M., PAI, G., VAN AKEN, S., UMAYAM, L., TALLON, L. J., GILL, J. E., ADAMS, M. D., CARRERA, A. J., CREASY, T. H., GOODMAN, H. M., SOMERVILLE, C. R., COPENHAVER, G. P., PREUSS, D., NIERMAN, W. C., WHITE, O., EISEN, J. A., SALZBERG, S. L., FRASER, C. M. & VENTER, J. C. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature,* 402**,** 761.

LIOTENBERG, S., NORTH, H. & MARION-POLL, A. 1999. Molecular biology and regulation of abscisic acid biosynthesis in plants. *Plant Physiology and Biochemistry,* 37**,** 341-350.

LIU, L., SHAO, Z., ZHANG, M. & WANG, Q. 2015. Regulation of carotenoid metabolism in tomato. *Molecular Plant,* 8**,** 28-39.

LIU, Q., ZHANG, W., WANG, H., LI, Y., LIU, W., WANG, Q., LIU, D., CHEN, N. & JIANG, W. 2016. Validation of a HILIC method for the analysis of ergothioneine in fermentation broth. *Journal of Chromatographic Science,* 54**,** 934-938.

LIU, R. & DICKERSON, J. 2017. Strawberry: Fast and accurate genome-guided transcript reconstruction and quantification from RNA-Seq. *PLoS Computational Biology,* 13**,** e1005851.

LIU, Z. & ROCHFORT, S. 2013. A fast liquid chromatography–mass spectrometry (LC–MS) method for quantification of major polar metabolites in plants. *Journal of Chromatography B,* 912**,** 8-15.

LOW, F. C. & BONNER, J. 1985. Characterisation of the nuclear genome of *Hevea brasiliensis*. *In:* RUBBER RESEARCH INSTITUTE OF MALAYSIA (ed.) *Proceedings of the International Rubber Conference.* Kuala Lumpur: Rubber Research Institute of Malaysia.

LU, S., VAN ECK, J., ZHOU, X., LOPEZ, A. B., O'HALLORAN, D. M., COSMAN, K. M., CONLIN, B. J., PAOLILLO, D. J., GARVIN, D. F., VREBALOV, J., KOCHIAN, L. V., KÜPPER, H., EARLE, E. D., CAO, J. & LI, L. 2006. The cauliflower *or* gene encodes a dnaJ cysteine-rich domain-

containing protein that mediates high levels of β-carotene accumulation. *The Plant Cell,* 18**,** 3594-3605.

LUNG, S.-C. C. & LIU, C.-H. 2015. Fast analysis of 29 polycyclic aromatic hydrocarbons (PAHs) and nitro-PAHs with ultra-high performance liquid chromatography-atmospheric pressure photoionization-tandem mass spectrometry. *Scientific Reports,* 5**,** 12992.

LYNEN, F. 1967. Biosynthesis pathways from acetate to natural products. Activity of the enzymes in rubber synthesis. *Pure and Applied Chemistry,* 14**,** 137-167.

LYNEN, F. 1969. Biochemical problems of rubber biosynthesis. *Journal of Rubber Research Institute of Malaya,* 21**,** 389-406.

MAKITA, Y., NG, K. K., VEERA SINGHAM, G., KAWASHIMA, M., HIRAKAWA, H., SATO, S., OTHMAN, A. S. & MATSUI, M. 2017. Large-scale collection of full-length cDNA and transcriptome analysis in *Hevea brasiliensis*. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes,* 24**,** 159-167.

MALAYSIAN RUBBER BOARD 2009. *Rubber Plantation and Processing Technologies,* Kuala Lumpur, Malaysia, Malaysian Rubber Board.

MANLEY, L. J., MA, D. & LEVINE, S. S. 2016. Monitoring error rates in illumina sequencing. *Journal of Biomolecular Techniques : JBT,* 27**,** 125-128.

MANN, C. E. T. 1934. Notes on the yield performance of budded trees of proved clones in commercial practice. *Journal of Rubber Research Institute of Malaya,* 5**,** 139-144.

MANTELLO, C. C., CARDOSO-SILVA, C. B., DA SILVA, C. C., DE SOUZA, L. M., SCALOPPI JUNIOR, E. J., GONC¸ALVES, P. D. S., VICENTINI, R. & DE SOUZA, A. P. 2014. De novo assembly and trannscriptome analysis of the rubber tree (*Hevea brasiliensis*) and SNP markers development for rubber biosythesis. *PLoS ONE,* 9**,** e102665. doi:10.1371/journal.pone.0102665.

MARDIS, E., MCPHERSON, J., MARTIENSSEN, R., WILSON, R. K. & MCCOMBIE, W. R. 2002. What is finished, and why does it matter. *Genome Research,* 12**,** 669-671.

MARIN, E., NUSSAUME, L., QUESADA, A., GONNEAU, M., SOTTA, B., HUGUENEY, P., FREY, A. & MARION-POLL, A. 1996. Molecular identification of zeaxanthin epoxidase of Nicotiana plumbaginifolia, a gene involved in abscisic acid biosynthesis and corresponding to the ABA locus of *Arabidopsis thaliana*. *The EMBO Journal,* 15**,** 2331-2342.

MARKLEY, J. L., BRÜSCHWEILER, R., EDISON, A. S., EGHBALNIA, H. R., POWERS, R., RAFTERY, D. & WISHART, D. S. 2017. The future of NMR-based metabolomics. *Current Opinion in Biotechnology,* 43**,** 34-40.

MARQUEZ, Y., BROWN, J. W. S., SIMPSON, C., BARTA, A. & KALYNA, M. 2012. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Research,* 22**,** 1184-1195.

MARSHALL, A. G. & HENDRICKSON, C. L. 2002. Fourier transform ion cyclotron resonance detection: principles and experimental configurations. *International Journal of Mass Spectrometry,* 215**,** 59-75.

MASCHER, M., GUNDLACH, H., HIMMELBACH, A., BEIER, S., TWARDZIOK, S. O., WICKER, T., RADCHUK, V., DOCKTER, C., HEDLEY, P. E., RUSSELL, J., BAYER, M., RAMSAY, L., LIU, H., HABERER, G., ZHANG, X.-Q., ZHANG, Q., BARRERO, R. A., LI, L., TAUDIEN, S., GROTH, M., FELDER, M., HASTIE, A., ŠIMKOVÁ, H., STAŇKOVÁ, H., VRÁNA, J., CHAN, S., MUÑOZ-AMATRIAÍN, M., OUNIT, R., WANAMAKER, S., BOLSER, D., COLMSEE, C., SCHMUTZER, T., ALIYEVA-SCHNORR, L., GRASSO, S., TANSKANEN, J., CHAILYAN, A., SAMPATH, D., HEAVENS, D., CLISSOLD, L., CAO, S., CHAPMAN, B., DAI, F., HAN, Y., LI, H., LI, X., LIN, C., MCCOOKE, J. K., TAN, C., WANG, P., WANG, S., YIN, S., ZHOU, G., POLAND, J. A., BELLGARD, M. I., BORISJUK, L., HOUBEN, A., DOLEŽEL, J., AYLING, S., LONARDI, S., KERSEY, P., LANGRIDGE, P., MUEHLBAUER, G. J., CLARK, M. D., CACCAMO, M., SCHULMAN, A. H., MAYER, K. F. X., PLATZER, M., CLOSE, T. J., SCHOLZ, U., HANSSON, M., ZHANG, G., BRAUMANN, I., SPANNAGL, M., LI, C., WAUGH, R. & STEIN, N. 2017. A chromosome conformation capture ordered sequence of the barley genome. *Nature,* 544**,** 427.

MATSUMOTO, H., IKOMA, Y., KATO, M., KUNIGA, T., NAOKO, N. & YOSHIDA, T. 2007. Quantification of carotenoids in citrus fruits by LC-MS and comparison of patterns of seasonal changes for carotenoids among citrus varieties. *J. Agric. Food Chem.,* 55**,** 2356-2368.

MCMULLEN, A. I. & MCSWEENEY, G. P. 1966. The biosynthesis of rubber: Incorporation of isopentenyl pyrophosphate into purified rubber particles by a soluble latex serum enzyme. *Biochemical Journal,* 101**,** 42-47.

MENDOZA-POUDEREUX, I., KUTZNER, E., HUBER, C., SEGURA, J., EISENREICH, W. & ARRILLAGA, I. 2015. Metabolic cross-talk between pathways of terpenoid backbone biosynthesis in spike lavender. *Plant Physiology and Biochemistry,* 95**,** 113-120.

MIĘKUS, N., KONIECZNA, L., KOWIAŃSKI, P., MORYŚ, J. & BĄCZEK, T. 2017. HILIC-MS rat brain analysis, a new approach for the study of ischemic attack. *Translational Neuroscience,* 8**,** 70-75.

MIGHELL, A. J., SMITH, N. R., ROBINSON, P. A. & MARKHAM, A. F. 2000. Vertebrate pseudogenes. *FEBS Letters,* 468**,** 109-114.

MILNE, I., LINDNER, D., BAYER, M., HUSMEIER, D., MCGUIRE, G., MARSHALL, D. F. & WRIGHT, F. 2009. TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics,* 25**,** 126-127.

MILNE, I., SHAW, P., STEPHEN, G., BAYER, M., CARDLE, L., THOMAS, W. T. B., FLAVELL, A. J. & MARSHALL, D. 2010. Flapjack—graphical genotype visualization. *Bioinformatics,* 26**,** 3133-3134.

MILNE, I., STEPHEN, G., BAYER, M., COCK, P. J. A., PRITCHARD, L., CARDLE, L., SHAW, P. D. & MARSHALL, D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics,* 14**,** 193-202.

MOHAMED-SATHIK, M. B., LUKE, L. P., RAJAMANI, A., KURUVILLA, L., SUMESH, K. V. & THOMAS, M. 2018. De novo transcriptome analysis of abiotic stress-responsive transcripts of *Hevea brasiliensis*. *Molecular Breeding,* 38**,** 32.

MOIR, G. F. J. 1959. Ultracentrifugation and staining of *Hevea* latex. *Nature,* `84**,** 1626.

MOLLISON, E. M. B., CLAVIJO, B. J., ACCINELLI, G. G., ANISSIN, M., SITI-ARIJA, M.-A., ATAN, S., BAHARI, A., BAKER, D., BARKER, T., BENNETT, G., CAIM, S., CHOW, K.-S., CUGLIANDOLO, F., DI COLA, A., DROU, N., HEAVENS, D., KHALIQ, I., MCLAY, K., PIDDOCK, R., RAMIREZ, R., RIBEIRO, C., SERKIOVA, V., SWARBECK, D., THOMPSON, C. E., WAITE, D., WARNER, D., WATKINS, C., CACCAMO, M. & KOLESNIKOVA-ALLEN, M. A. 2014. Harnessing the future natural rubber supply - *Hevea brasiliensis* genome sequence as a foundation for targeted crop improvement. *Plant Genomics Congress.* London, UK.

MONTORO, P., GRAMDI, S., KUSWANHADI, LEFRANÇOIS, C., NEMROD, G., ARGOUT, X., BAURENS, F.-C., LECLERCQ, J., RIO, M. & SABAU, X. 2008. Ethylene-regulated genes in *Hevea brasiliensis*: effect of ethylene and wounding in young budded plants of three clones with contrasting metabolisms *In:* IRRDB (ed.) *IRRDB Natural Rubber Conference.* Kuala Lumpur, Malaisie: s.n.

MORRIS, M. & LAKIN, B. 1995. Natural rubber latex production, composition and specifications. *Rubber Division, American Chemical Society Educational Symposium on Natural Rubber.* Philadelphia, Pennsylvania.

MORRIS, W. L., DUCREUX, L., GRIFFITHS, D. W., STEWART, D., DAVIES, H. V. & TAYLOR, M. A. 2004. Carotenogenesis during tuber development and storage in potato. *Journal of Experimental Botany,* 55**,** 975-982.

MURPHY, D. J. & VANCE, J. 1999. Mechanisms of lipid-body formation. *Trends in Biochemical Sciences,* 24**,** 109-115.

NAWAMAWAT, K., SAKDAPIPANICH, J. T., HO, C. C., MA, Y., SONG, J. & VANSCO, J. G. 2011. Surface nanostructure of *Hevea* brasiliensis natural rubber latex particles *Colloids and Surfaces A: Physicochemical and Engineering Aspects,* 390**,** 157-166.

NAZ, S., GALLART-AYALA, H., REINKE, S. N., MATHON, C., BLANKLEY, R., CHALECKIS, R. & WHEELOCK, C. E. 2017. Development of a liquid chromatography–high resolution mass spectrometry metabolomics method with high specificity for metabolite identification using all ion fragmentation acquisition. *Analytical Chemistry,* 89**,** 7933-7942.

NICA, A. C., ONGEN, H., IRMINGER, J.-C., BOSCO, D., BERNEY, T., ANTONARAKIS, S. E., HALBAN, P. A. & DERMITZAKIS, E. T. 2013. Cell-type, allelic, and genetic signatures in the human pancreatic beta cell transcriptome. *Genome Research,* 23**,** 1554-1562.

NISAR, N., LI, L., LU, S., KHIN, NAY C. & POGSON, BARRY J. 2015. Carotenoid metabolism in plants. *Molecular Plant,* 8**,** 68-82.

NÜRENBERG, G. & VOLMER, D. A. 2012. The analytical determination of isoprenoid intermediates from the mevalonate pathway. *Analytical and Bioanalytical Chemistry,* 402**,** 671-685.

NURMI-ROHAYU, A. M. 2017. *RE: Large Scale Clone Trials for RRIM2000 and RRIM 3000 series. Personal Communication.*

OH, S. K., KANG, H., SHIN, D. H., YANG, J., CHOW, K.-S., YEANG, H. Y., WAGNER, B., BREITENEDER, H. & HAN, K.-H. 1999. Isolation, characterization, and functional analysis of a novel cDNA clone encoding a small rubber particle protein from *Hevea brasiliensis*. *Journal of Biological Chemistry,* 274**,** 17132-17138.

OHYA, N., TANAKA, Y., WITITSUWAANNAKUL, D. & KOYAMA, T. 2000. Activity of rubber transferase and rubber particle size in *Hevea* latex. *Journal of Rubber Research,* 3**,** 214-221.

ONG, E. L. 2000. Characterisation of new latex-timber clones of natural rubber. *Journal of Applied Polymer Science,* 78**,** 1517-1521.

ONG, S., RAMLI, O., MD-ZAIN, A.-A., OTHMAN, H., MASAHULING, B. & MOHD-NOOR, A.-G. 1995. RRIM Planting Recommendations 1995-1997. *Rubber Growers' Conference.* Kuala Lumpur.

ONG, S. H., OTHMAN, R., OTHMAN, H., BENONG, M. & NAIMAH, I. 1994. Rubber breeding, progress and strategies to meet future needs of the plantation industry. *International Planters' Conference.* Kuala Lumpur: Rubber Research Institute of Malaysia.

PAKIANATHAN, S. W., SAMSIDAR, H., SIVAKUMAR, S. & GOMEZ, J. B. 1982. Physiological and anatomical investigations on long-term ethephon-stimulated trees. *Journal of Rubber Research Institute of Malaysia,* 30**,** 63-79.

PAOLILLO, D. J., GARVIN, D. F. & PARTHASARATHY, M. V. 2004. The chromoplasts of Or mutants of cauliflower (*Brassica oleracea* L. var. botrytis). *Protoplasma,* 224**,** 245-253.

PARK, S.-Y., FUNG, P., NISHIMURA, N., JENSEN, D. R., FUJII, H., ZHAO, Y., LUMBA, S., SANTIAGO, J., RODRIGUES, A., CHOW, T.-F. F., ALFRED, S. E., BONETTA, D., FINKELSTEIN, R., PROVART, N. J., DESVEAUX, D., RODRIGUEZ, P. L., MCCOURT, P., ZHU, J.-K., SCHROEDER, J. I., VOLKMAN, B. F. & CUTLER, S. R. 2009. Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins. *Science,* 324**,** 1068-1071.

PATRO, R., DUGGAL, G., LOVE, M. I., IRIZARRY, R. A. & KINGSFORD, C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods,* 14**,** 417.

PERTEA, M., PERTEA, G. M., ANTONESCU, C. M., CHANG, T.-C., MENDELL, J. T. & SALZBERG, S. L. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology,* 33**,** 290-295.

PHATTHIYA, A., TAKAHASHI, S., CHAREONTHIPHAKORN, N., KOYAMA, T., WITITSUWANNAKUL, D. & WITITSUWANNAKUL, R. 2007. Cloning and expression of the gene encoding solanesyl diphosphate synthase from *Hevea brasiliensis*. *Plant Science,* 172**,** 824-831.

PIRRELLO, J., LECLERCQ, J., DESSAILLY, F., RIO, M., PIYATRAKUL, P., KUSWANHADI, K., TANG, C. & MONTORO, P. 2014. Transcriptional and post-transcriptional regulation of the jasmonate signalling pathway in response to abiotic and harvesting stress in *Hevea brasiliensis*. *BMC Plant Biology,* 14**,** 341.

PONCIANO, G., MCMAHAN, C. M., XIE, W., LAZO, G. R., COFFELT, T. A., COLLINS-SILVA, J., NURAL-TABAN, A., GOLLERY, M., SHINTANI, D. K. & WHALEN, M. C. 2012.

Transcriptome and gene expression analysis in cold-acclimated guayule (*Phartenium argentatum*) rubber-producing tissue. *Phytochemistry,* 79**,** 57-66.

POOTAKHAM, W., RUANG-AREERATE, P., JOMCHAI, N., SONTHIROD, C., SANGSRAKRU, D., YOOCHA, T., THEERAWATTANASUK, K., NIRAPATHPONGPORN, K., ROMRUENSUKHAROM, P., TRAGOONRUNG, S. & TANGPHATSORNRUANG, S. 2015. Construction of a high-density integrated genetic linkage map of rubber tree (*Hevea brasiliensis*) using genotyping-by-sequencing (GBS). *Frontiers in Plant Science,* 6**,** 367.

POOTAKHAM, W., SONTHIROD, C., NAKTANG, C., RUANG-AREERATE, P., YOOCHA, T., SANGSRAKRU, D., THEERAWATTANASUK, K., RATTANAWONG, R., LEKAWIPAT, N. & TANGPHATSORNRUANG, S. 2017. De novo hybrid assembly of the rubber tree genome reveals evidence of paleotetraploidy in *Hevea* species. *Scientific Reports,* 7**,** 41457.

PRIYA, P., VENKATACHALAM, P. & THULASEEDHARAN, A. 2007. Differential expression pattern of rubber elongation factor (REF) mRNA transcripts from high and low yielding clones of rubber tree (*Hevea brasiliensis* Muell. Arg.). *Plant Cell Reports,* 26**,** 1833-1838.

PRIYADARSHAN, P. M. 2017a. Plant Structure and Ecophysiology. *In:* PRIYADARSHAN, P. M. (ed.) *Biology of Hevea Rubber.* Cham: Springer International Publishing.

PRIYADARSHAN, P. M. 2017b. Refinements to *Hevea* rubber breeding. *Tree Genetics & Genomes,* 13**,** 20.

PUSKAS, J. E., GAUTRIAUD, E., DEFFIEUX, A. & KENNEDY, J. P. 2006. Natural rubber biosynthesis—A living carbocationic polymerization? *Progress in Polymer Science,* 31**,** 533-548.

PÜTTER, K. M., VAN DEENEN, N., UNLAND, K., PRÜFER, D. & SCHULZE GRONOVER, C. 2017. Isoprenoid biosynthesis in dandelion latex is enhanced by the overexpression of three key enzymes involved in the mevalonate pathway. *BMC Plant Biology,* 17**,** 88.

RAHMAN, A. Y. A., USHARAJ, A. O., MISRA, B. B., THOTTATHIL, G. P., JAYASEKARANA, K., FENG, Y., HOU, S., ONG, S. Y., NG, F. L., LEE, L. S., TAN, H. S., SAKAFF, M. K., TEH, B. S., KHOO, B. F., BADAI, S. S., AB-AZIZ, N., YURYEV, A., KNUDSEN, B., DIONNE-LAPORTE, A., MCHUNU, N. P., YU, Q., LANGSTON, B. J., FREITAS, T. A., YOUNG, A. G., CHEN, R., WANG, L., NAJIMUDIN, N., SAITO, J. A. & ALAM, M. 2013. Draft genome sequence of the rubber tree *Hevea* brasiliensis. *BMC Genomics,* 14**,** doi:10.1186/1471-2164-14-75.

RAMLI, O., MASAHULING, B., MD-ZAIN, A. A., ZARAWI, A. G., A, M.-N. M., ADAM-MALIK, A. Z. & W, O. C. 2005. Performance of RRIM2000 series clones in large scale clone trials and monitored development project. *Proceeding of Rubber Planters' Conference.* Kuala Lumpur: Malaysian Rubber Board.

RAO, G. P., SUMA, K., MADHAVAN, J. & VARGHESE, Y. A. 2013. Variability in Wild Germplasm of Natural Rubber (*Hevea brasiliensis* Muell. Arg.). *Silvae Genetica.*

RATNAM, W., CHOONG, C. Y. & JAVED, M. A. 2017. Development of Genomic Resources and Assessing Their Potential for Accelerated Acacia Breeding. *In:* ABDULLAH, S. N. A., CHAI-LING, H. & WAGSTAFF, C. (eds.) *Crop Improvement: Sustainability Through Leading-Edge Technology.* Cham: Springer International Publishing.

REYES-CHIN-WO, S., WANG, Z., YANG, X., KOZIK, A., ARIKIT, S., SONG, C., XIA, L., FROENICKE, L., LAVELLE, D. O., TRUCO, M.-J., XIA, R., ZHU, S., XU, C., XU, H., XU, X., COX, K., KORF, I., MEYERS, B. C. & MICHELMORE, R. W. 2017. Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature Communications,* 8**,** 14953.

RICHES, J. P. & GOODING, E. G. B. 1952. Studies in the physiology of latex. I. Latex flow on tapping -theoretical considerations. *New Phytologist,* 51**,** 1-10.

RICHTERICH, P. 1998. Estimation of Errors in "Raw" DNA Sequences: A Validation Study. *Genome Research,* 8**,** 251-259.

RIVERA, S. M., CHRISTOU, P. & CANELA-GARAYOA, R. 2014. Identification of carotenoids using mass spectrometry. *Mass Spectrometry Reviews,* 33**,** 353-372.

ROBINSON, M. D., MCCARTHY, D. J. & SMYTH, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26.

ROBINSON, M. D. & OSHLACK, A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol,* 11.

ROHMER, M. 1999. The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants[dagger]. *Natural Product Reports,* 16**,** 565-574.

ROUNSLEY, S. D., GLODEK, A., SUTTON, G., ADAMS, M. D., SOMERVILLE, C. R., VENTER, J. C. & KERLAVAGE, A. R. 1996. The construction of Arabidopsis expressed sequence tag assemblies (a new resource to facilitate gene identification). *Plant Physiology,* 112**,** 1177-1183.

RUBBER RESEARCH INSTITUTE OF MALAYSIA 1969. Planting Recommendations 1969-1970. *Planters' Bulletin.* Kuala Lumpur: Rubber Research Institute of Malaysia.

RUBBER RESEARCH INSTITUTE OF MALAYSIA 1970. Review of Modern Clones: RRIM 600. *Planters' Bulletin,* 107**,** 49-64.

RUBBER RESEARCH INSTITUTE OF MALAYSIA 1973. Methods for measuring the dry rubber content of field latex *In:* RUBBER RESEARCH INSTITUTE OF MALAYSIA (ed.) *Planters' Bulletin.* Kuala Lumpur: Rubber Research Institute of Malaysia.

RUIZ-SOLA, M. Á. & RODRÍGUEZ-CONCEPCIÓN, M. 2012. Carotenoid Biosynthesis in Arabidopsis: A Colorful Pathway. *The Arabidopsis Book / American Society of Plant Biologists,* 10**,** e0158.

RUIZ, D., EGEA, J., TOMÁS-BARBERÁN, F. A. & GIL, M. I. 2005. Carotenoids from new apricot (*Prunus armeniaca* l.) varieties and their relationship with flesh and skin color. *Journal of Agricultural and Food Chemistry,* 53**,** 6368-6374.

SAFIAH, A. 2012. *Construction of scaffold genetic linkage maps for two Hevea families: PB 5/51 X IAN 873 and RRIM 937 X RRIM 600*.

SAJARI, R., ABDUL RAZAK, N. H., YUSOF, F., ARIF, S. A. M., PERKINS, M. & YEANG, H. Y. 2014. Improved efficiency of tocotrienol extraction from fresh and processed latex. *Journal of Rubber Research,* 17**,** 245-260.

SAKDAPIPANICH, J. T. 2006. Composition of color substances of *Hevea brasiliensis* natural rubber. *Rubber Chemistry and Technology,* 80**,** 212-230.

SAKDAPIPANICH, J. T. 2007. Structural characterisation of natural rubber based on recent evidence from selective enzymatic treatments. *Journal of Bioscience and Bioengineering,* 103**,** 287-292.

SAKDAPIPANICH, J. T., SUKSUJARITPORN, S. & TANAKA, Y. 1999. Structural characterisation of the small rubber particles in fresh *Hevea* latex. *Journal of Rubber Research,* 2**,** 160-168.

SALAMOV, A. A. & SOLOVYEV, V. V. 2000. Ab initio Gene Finding in Drosophila Genomic DNA. *Genome Research,* 10**,** 516-522.

SALEM, M. A., JÜPPNER, J., BAJDZIENKO, K. & GIAVALISCO, P. 2016. Protocol: a fast, comprehensive and reproducible one-step extraction method for the rapid preparation of polar and semi-polar metabolites, lipids, proteins, starch and cell wall polymers from a single sample. *Plant Methods,* 12**,** 45.

SALGADO, L. R., KOOP, D. M., PINHEIRO, D. G., RIVALLAN, R., LE GUEN, V., NICOLÁS, M. F., DE ALMEIDA, L. G. P., ROCHA, V. R., MAGALHÃES, M., GERBER, A. L., FIGUEIRA, A., CASCARDO, J. C. D. M., DE VASCONCELOS, A. R., SILVA, W. A., COUTINHO, L. L. & GARCIA, D. 2014. De novo transcriptome analysis of *Hevea brasiliensis* tissues by RNA-seq and screening for molecular markers. *BMC Genomics,* 15**,** 236.

SALMELA, L. & RIVALS, E. 2014. LoRDEC: accurate and efficient long read error correction. *Bioinformatics,* 30**,** 3506-3514.

SALOMEZ, M., SUBILEAU, M., INTAPUN, J., BONFILS, F., SAINTE-BEUVE, J., VAYSSE, L. & DUBREUCQ, E. 2014. MIcro-organisms in latex and natural rubber coagula of *Hevea brasiliensis* and their impact on rubber composition structure and properties. *Journal of Applied Microbiology,* 117**,** 921-929.

SANDO, T., HAYASHI, T., TAKEDA, T., AKIYAMA, Y., NAKAZAWA, Y., FUKUSAKI, E. & KOBAYASHI, A. 2009. Histochemical study of detailed laticifer structure and rubber biosynthesis-

related protein localization in *Hevea brasiliensis* using spectral confocal laser scanning microscopy. *Planta,* 230**,** 215-225.

SANDO, T., TAKAOKA, C., MUKAI, Y., YAMASHITA, A., HATTORI, M., OGASAWARA, N., FUKUSAKI, E. & KOBAYASHI, A. 2008a. Cloning and characterization of mevalonate pathway genes in a natural rubber producing plant, *Hevea brasiliensis. Bioscience, biotechnology, and biochemistry,* 72**,** 2049-2060.

SANDO, T., TAKENO, S., WATANABE, N., OKUMOTO, H., KUZUYAMA, T., YAMASHITA, A., HATTORI, M., OGASAWARA, N., FUKUSAKI, E. & KOBAYASHI, A. 2008b. Cloning and characterization of the 2- c -methyl- d -erythritol 4-phosphate (MEP) pathway genes of a natural-rubber producing plant, *Hevea brasiliensis*. *Bioscience, Biotechnology, and Biochemistry,* 72**,** 2903-2917.

SAZONOVS, A. & BARRETT, J. C. 2018. Rare-variant studies to complement genome-wide association studies. *Annual Review of Genomics and Human Genetics,* 19**,** 97-112.

SCHIEDT, K. & LIAAN-JENSEN, S. 1995. Isolation and analysis. *In:* BRITTON, G., LIAAN-JENSEN, S. & PFANDER, H. (eds.) *Carotenoids.* Basel, Switzerland: Birkhäuser Verlag.

SCHIRMER, M., D'AMORE, R., IJAZ, U. Z., HALL, N. & QUINCE, C. 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics,* 17**,** 125.

SCHMIDT, T., HILLEBRAND, A., WURBS, D., WAHLER, D., LENDERS, M., GRONOVER, C. S. & PRÜFER, D. 2010a. Molecular cloning and characterization of rubber biosynthetic genes from *Taraxacum koksaghyz*. *Plant Molecular Biology,* 28**,** 277-284.

SCHMIDT, T., LENDERS, M., HILLEBRAND, A., VAN DEENEN, N., MUNT, O., REICHEL, R., EISENREICH, W., FISCHER, R., PRÜFER, D. & GRONOVER, C. S. 2010b. Characterisation of rubber particles and rubber chain elongation in *Taraxacum koksaghyz*. *BMC Biochemistry,* 11**,** doi:10.1186/1471-2091-11-11.

SCHROEDER, A., MUELLER, O., STOCKER, S., SALOWSKY, R., LEIBER, M. & GASSMANN, M. 2006. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol,* 7.

SEDLAZECK, F. J., LEE, H., DARBY, C. A. & SCHATZ, M. C. 2018. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics,* 19**,** 329-346.

SHAW, P. D., GRAHAM, M., KENNEDY, J., MILNE, I. & MARSHALL, D. F. 2014. Helium: visualization of large scale plant pedigrees. *BMC Bioinformatics,* 15**,** 259.

SHERIEF, P. M. & SETHURAJ, M. R. 1978. The role of lipids and proteins in the mechanism of latex vessel plugging in *Hevea brasiliensis*. *Physiologia Plantarum,* 42**,** 351-353.

SHEWMAKER, C. K., SHEEHY, J. A., DALEY, M., COLBURN, S. & KE, D. Y. 1999. Seed-specific overexpression of phytoene synthase: increase in carotenoids and other metabolic effects. *The Plant Journal,* 20**,** 401-412.

SHIMIZU, T., TANIZAWA, Y., MOCHIZUKI, T., NAGASAKI, H., YOSHIOKA, T., TOYODA, A., FUJIYAMA, A., KAMINUMA, E. & NAKAMURA, Y. 2017. Draft sequencing of the heterozygous diploid genome of satsuma (*Citrus unshiu* Marc.) USING A HYBRID ASSEMBLY APPROACH. *Frontiers in Genetics,* 8.

SIMÃO, F. A., WATERHOUSE, R. M., IOANNIDIS, P., KRIVENTSEVA, E. V. & ZDOBNOV, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics,* 31**,** 3210-3212.

SIMMONDS, N. W. The strategy of rubber breeding. Proceedings of the International Rubber Conference, 1985 Kuala Lumpur. Rubber Research Institute of Malaysia, 115-126.

SINGH, A. P., WI, S. D., CHUNG, G. C., KIM, Y. S. & KANG, H. 2003. The micromorphology and protein characterisation of rubber particles in *Ficus carica, Ficus benghalensis* and *Hevea brasiliensis*. *Journal of Experimental Botany,* 54**,** 985-992.

SMITH-UNNA, R., BOURSNELL, C., PATRO, R., HIBBERD, J. M. & KELLY, S. 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Research,* 26**,** 1134-1144.

SONESON, C., LOVE, M. I., ROBINSON, M. D., SONESON, C., LOVE, M. I. & ROBINSON, M. D. 2016. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research,* 4**,** 1521.

SOOKMARK, U., PUJADE-RENAUD, V., CHRESTIN, H., LACOTE, R., NAIYANETR, C., SEGUIN, M., ROMRUENSUKHAROM, P. & NARANGAJAVANA, J. 2002. Characterization of polypeptides accumulated in the latex cytosol of rubber trees affected by the tapping panel dryness syndrome. *Plant and Cell Physiology,* 43**,** 1323-1333.

SORATANA, K., RASUTIS, D., AZARABADI, H., ERANKI, P. L. & LANDIS, A. E. 2017. Guayule as an alternative source of natural rubber: A comparative life cycle assessment with *Hevea* and synthetic rubber. *Journal of Cleaner Production,* 159**,** 271-280.

SOUTHORN, W. A. & YIP, E. 1968. Latex flow studies III. Electrostatic considerations in the colloidal stability of fresh *Hevea* latex. *Journal of Rubber Research Institute of Malaya,* 20**,** 201-215.

SPRIGGS, A., HENDERSON, S. T., HAND, M. L., JOHNSON, S. D., TAYLOR, J. M. & KOLTUNOW, A. 2018. Assembled genomic and tissue-specific transcriptomic data resources for two genetically distinct lines of Cowpea (*Vigna unguiculata* (L.) Walp). *Gates Open Research,* 2**,** 7.

STANKE, M., DIEKHANS, M., BAERTSCH, R. & HAUSSLER, D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics,* 24**,** 637-644.

STEINBÜCHEL, A. 2003. Production of rubber-like polymers by microorganisms. *Current Opinion in Microbiology,* 6**,** 261-270.

STEWART, I. & WHEATON, T. 1973. Conversion of β-apo-8'-carotenal to cintranaxanthin during the isolations of carotenoids from Citrus. *Phytochemistry,* 12**,** 2947-2951.

SUBRAMANIAM, A. 1980. Molecular weight and molecular weight distribution of natural rubber. *Rubber Research Institute of Malaysia Technology Bulletin.* Rubber Research Institute of Malaysia.

SUBROTO, T., VAN KONINGSVELD, G. A., SCHREUDER, H. A., SOEDJANAATMADJA, U. M. S. & BEINTEMA, J. J. 1996. Chitinase and β-1,3-glucanase in the lutoid-body fraction of *Hevea* latex. *Phytochemistry,* 43**,** 29-37.

SUN, T., YUAN, H., CAO, H., YAZDANI, M., TADMOR, Y. & LI, L. 2018. Carotenoid metabolism in plants: The role of plastids. *Molecular Plant,* 11**,** 58-74.

SUNDERASAN, E., LING, S. K., ARIF, S. A. M. & YEANG, H. Y. 2007. Deciphering rubber particle destabilisation by *Hevea* bark extract. *Journal of Rubber Research,* 10**,** 97-106.

SUWANMANEE, P., SIRINUPONG, N. & SUVACHITTANONT, W. 2013. Regulation of 3-Hydroxy-3-Methylglutaryl-CoA Synthase and 3-Hydroxy-3-Methylglutaryl-CoA Reductase and Rubber Biosynthesis of *Hevea brasiliensis* (B.H.K.) Mull. Arg. *In:* BACH, T. J. & ROHMER, M. (eds.) *Isoprenoid Synthesis in Plants and Microorganisms: New Concepts and Experimental Approaches.* New York, NY: Springer New York.

TAN, D., HU, X., FU, L., KUMPEANGKEAW, A., DING, Z., SUN, X. & ZHANG, J. 2017. Comparative morphology and transcriptome analysis reveals distinct functions of the primary and secondary laticifer cells in the rubber tree. *Scientific Reports,* 7**,** 3126.

TAN, H. 1978. Assessment of parental performance for yield in *Hevea* breeding. *Euphytica,* 27**,** 521-528.

TANAKA, M., ISHIGURO, K., OKI, T. & OKUNO, S. 2017. Functional components in sweetpotato and their genetic improvement. *Breeding Science,* 67**,** 52-61.

TANAKA, Y. 1985. Structural characterisation of cis-polyisoprenes from sunflower, *Hevea* and guayule. *Proceedings of the International Rubber Conference* Kuala Lumpur.

TANAKA, Y., ENG, A. H., OHYA, N., NISHIYAMA, N., TANGPAKDEE, J., KAWAHARA, S. & WITITSUWANNAKUL, R. 1996. Initiation of rubber biosynthesis in *Hevea brasiliensis:* characterisation of initiating species by structural analysis. *Phytochemistry,* 41**,** 1501-1505.

TANG, C., YANG, M., FANG, Y., LUO, Y., GAO, S., XIAO, X., AN, Z., ZHOU, B., ZHANG, B., TAN, X., YEANG, H.-Y., QIN, Y., YANG, J., LIN, Q., MEI, H., MONTORO, P., LONG, X., QI, J., HUA, Y., HE, Z., SUN, M., LI, W., ZENG, X., CHENG, H., LIU, Y., YANG, J., TIAN, W., ZHUANG, N., ZENG, R., LI, D., HE, P., LI, Z., ZOU, Z., LI, S., LI, C., WANG, J., WEI, D., LAI, C.-Q., LUO, W., YU, J., HU, S. & HUANG, H. 2016. The rubber tree genome reveals new insights into rubber production and species adaptation. *Nature Plants,* 2**,** 16073.

TARDAGUILA, M., DE LA FUENTE, L., MARTI, C., PEREIRA, C., DEL RISCO, H., FERRELL, M., MELLADO, M., MACCHIETTO, M., VERHEGGEN, K., EDELMANN, M., EZKURDIA, I., VAZQUEZ, J., TRESS, M., MORTAZAVI, A., MARTENS, L., RODRIGUEZ-NAVARRO, S., MORENO, V. & CONESA, A. 2017. SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification. *bioRxiv*.

TATEYAMA, S., WITITSUWANNAKUL, R., WITITSUWANNAKUL, D., SAGAMI, H. & OGURA, K. 1999. Dolichols of rubber plant, ginkgo and pine. *Phytochemistry,* 51**,** 11-15.

THORUP, T. A., TANYOLAC, B., LIVINGSTONE, K. D., POPOVSKY, S., PARAN, I. & JAHN, M. 2000. Candidate gene analysis of organ pigmentation loci in the Solanaceae. *Proceedings of the National Academy of Sciences of the United States of America,* 97**,** 11192-11197.

THORVALDSDÓTTIR, H., ROBINSON, J. T. & MESIROV, J. P. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics,* 14**,** 178-192.

TOLSTIKOV, V. V. & FIEHN, O. 2002. Analysis of highly polar compounds of plant origin: Combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Analytical Biochemistry,* 301**,** 298-307.

TONG, Z., WANG, D., SUN, Y., YANG, Q., MENG, X., WANG, L., FENG, W., LI, L., WURTELE, E. & WANG, X. 2017. Comparative proteomics of rubber latex revealed multiple protein species of ref/srpp family respond diversely to ethylene stimulation among different rubber tree clones. *International Journal of Molecular Sciences,* 18**,** 958.

TØRRESEN, O. K., STAR, B., JENTOFT, S., REINAR, W. B., GROVE, H., MILLER, J. R., WALENZ, B. P., KNIGHT, J., EKHOLM, J. M., PELUSO, P., EDVARDSEN, R. B., TOOMING-KLUNDERUD, A., SKAGE, M., LIEN, S., JAKOBSEN, K. S. & NEDERBRAGT, A. J. 2017. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics,* 18**,** 95.

TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., SALZBERG, S. L., RINN, J. L. & PACHTER, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols,* 7**,** 562-578.

TSENG, E. 2016. Cogent: Reconstructing the coding region genome using full-length transcriptome sequences without a reference genome. *In:* BIOSCIENCES, P. (ed.). Menlo Park, CA, USA.

TSENG, E. 2017. *Cupcake ToFU: supporting scripts for Iso Seq after clustering step* [Online]. Available: https://github.com/Magdoll/cDNA_Cupcake/wiki/Cupcake-ToFU:-supporting-scripts-for-Iso-Seq-after-clustering-step [Accessed].

UDOH, L. I., GEDIL, M., PARKES, E. Y., KULAKOW, P., ADESOYE, A., NWUBA, C. & RABBI, I. Y. 2017. Candidate gene sequencing and validation of SNP markers linked to carotenoid content in cassava (*Manihot esculenta* Crantz). *Molecular Breeding,* 37**,** 123.

UYGUN, S., PENG, C., LEHTI-SHIU, M. D., LAST, R. L. & SHIU, S.-H. 2016. Utility and limitations of using gene expression data to identify functional associations. *PLOS Computational Biology,* 12**,** e1005244.

VAN BEILEN, J. B. & POIRIER, Y. 2007. Guayule and Russian dandelion as alternative sources of natural rubber. *Critical Reviews in Biotechnology,* 27**,** 217-231.

VENKATACHALAM, P., THULASEEDHARAN, A. & RAGHOTHAMA, K. 2007. Identification of expression profiles of tapping panel dryness (TPD) associated genes from the latex of rubber tree (*Hevea brasiliensis* Muell. Arg.). *Planta,* 226**,** 499-515.

VERHAAR, G. 1959. Natural latex as a colloidal system. *Rubber Chemistry and Technology,* 32**,** 1627-1659.

VRANOVÁ, E., COMAN, D. & GRUISSEM, W. 2013. Network analysis of the MVA and MEP pathways for isoprenoid synthesis. *Annual Review of Plant Biology,* 64**,** 665-700.

WADEESIRISAK, K., CASTANO, S., BERTHELOT, K., VAYSSE, L., BONFILS, F., PERUCH, F., RATTANAPORN, K., LIENGPRAYOON, S., LECOMTE, S. & BOTTIER, C. 2017. Rubber particle proteins REF1 and SRPP1 interact differently with native lipids extracted from *Hevea brasiliensis* latex. *Biochimica et Biophysica Acta (BBA) - Biomembranes,* 1859**,** 201-210.

WANG, B., TSENG, E., REGULSKI, M., CLARK, T. A., HON, T., JIAO, Y., LU, Z., OLSON, A., STEIN, J. C. & WARE, D. 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature Communications,* 7**,** 11708.

WANG, C., ZENG, J., LI, Y., HU, W., CHEN, L., MIAO, Y., DENG, P., YUAN, C., MA, C., ZANG, M., WANG, Q., LI, K., CHANG, J., WANG, Y., YANG, G. & HE, G. 2014. Enrichment of provitamin A content in wheat (*Triticum aestivum* L.) by introduction of the bacterial carotenoid biosynthetic genes CrtB and CrtI. *Journal of Experimental Biology,* 65**,** 2545-2556.

WANG, D., SUN, Y., CHANG, L., TONG, Z., XIE, Q., JIN, X., ZHU, L., HE, P., LI, H. & WANG, X. 2018. Subcellular proteome profiles of different latex fractions revealed washed solutions from rubber particles contain crucial enzymes for natural rubber biosynthesis. *Journal of Proteomics,* 182**,** 53-64.

WANG, X., SHI, M., WANG, D., CHEN, Y., CAI, F., ZHANG, S., WANG, L., TONG, Z. & TIAN, W.-M. 2013. Comparative proteomics of primary and secondary lutoids reveals that chitinase and glucanase play a crucial combined role in rubber particle aggregation in *Hevea brasiliensis*. *Journal of Proteome Research,* 12**,** 5146-5159.

WANG, Y., ZHAN, D.-F., LI, H.-L., GUO, D., ZHU, J.-H. & PENG, S.-Q. 2017. Transcriptome-wide identification and characterization of MYB transcription factor genes in the laticifer cells of *Hevea brasiliensis*. *Frontiers in Plant Science,* 8.

WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M. A., CLAMP, M. & BARTON, G. J. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25**,** 1189-1191.

WEEDON, B. C. L. & MOSS, G. P. 1995. Structure and nomenclature. *In:* LIAAN-JENSEN, S. & PFANDER, H. (eds.) *Carotenoids.* Basel, Switzerland: Birkhäuser Verlag.

WEIRATHER, J. L., DE CESARE, M., WANG, Y., PIAZZA, P., SEBASTIANO, V., WANG, X.-J., BUCK, D. & AU, K. F. 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research,* 6**,** 100.

WELSCH, R., ARANGO, J., BÄR, C., SALAZAR, B., AL-BABILI, S., BELTRÁN, J., CHAVARRIAGA, P., CEBALLOS, H., TOHME, J. & BEYER, P. 2010. Provitamin A Accumulation in Cassava (Manihot esculenta) Roots Driven by a Single Nucleotide Polymorphism in a Phytoene Synthase Gene. *The Plant Cell Online*.

WELSCH, R., WÜST, F., BÄR, C., AL-BABILI, S. & BEYER, P. 2008. A Third Phytoene Synthase Is Devoted to Abiotic Stress-Induced Abscisic Acid Formation in Rice and Defines Functional Diversification of Phytoene Synthase Genes. *Plant Physiology,* 147**,** 367-380.

WERNISCH, S. & PENNATHUR, S. 2016. Evaluation of coverage, retention patterns, and selectivity of seven liquid chromatographic methods for metabolomics. *Analytical and Bioanalytical Chemistry,* 408**,** 6079-6091.

WHEELER, N. C. & JECH, K. S. 1992. The use of electrophoretic markers in seed orchard research. *New Forests,* 6**,** 311-328.

WHITTLE, K. J., AUDLEY, B. G. & PENNOCK, J. F. 1967. The incorporation of [^{14}C]methionine into chromanols and quinones by *Hevea* brasiliensis latex *Biochemistry Journal,* 103**,** 21-22.

WILLIAMS, C. R., BACCARELLA, A., PARRISH, J. Z. & KIM, C. C. 2016. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics,* 17**,** 103.

WITITSUWAANNAKUL, D., RATTANAPITTAYAPORN, A., KOYAMA, T. & WITITSUWAANNAKUL, R. 2004. Involvement of *Hevea* latex organelle membrane proteins in the rubber biosynthesis activity and regulatory function. *Macromolecular Bioscience,* 4**,** 314-323.

WITITSUWANNAKUL, R., RUKSEREE, K., KANOKWIROON, K. & WITITSUWANNAKUL, D. 2008. A rubber particle protein specific for *Hevea* latex lectin binding involved in latex coagulation. *Phytochemistry,* 69**,** 1111-1118.

WOLSTENCROFT, K., HAINES, R., FELLOWS, D., WILLIAMS, A., WITHERS, D., OWEN, S., SOILAND-REYES, S., DUNLOP, I., NENADIC, A., FISHER, P., BHAGAT, J., BELHAJJAME, K., BACALL, F., HARDISTY, A., NIEVA DE LA HIDALGA, A., BALCAZAR VARGAS, M. P., SUFI, S. & GOBLE, C. 2013. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Research,* 41**,** W557-W561.

WÖLWER-RIECK, U., MAY, B., LANKES, C. & WÜST, M. 2014. Methylerythritol and mevalonate pathway contributions to biosynthesis of mono-, sesqui-, and diterpenes in glandular trichomes and leaves of *Stevia rebaudiana* Bertoni. *Journal of Agricultural and Food Chemistry,* 62**,** 2428-2435.

WONG, J. C., LAMBERT, R. J., WURTZEL, E. T. & ROCHEFORD, T. R. 2004. QTL and candidate genes phytoene synthase and ζ-carotene desaturase associated with the accumulation of carotenoids in maize. *Theoretical and Applied Genetics,* 108**,** 349-359.

WOO, C. H. 1973. Rubber coagulation by enzymes of *Hevea brasiliensis* latex *J. Rubb. Res. Inst. Malaysia,* 23**,** 323.

WRIGHT, L. P., ROHWER, J. M., GHIRARDO, A., HAMMERBACHER, A., ORTIZ-ALCAIDE, M., RAGUSCHKE, B., SCHNITZLER, J.-P., GERSHENZON, J. & PHILLIPS, M. A. 2014. Deoxyxylulose 5-phosphate synthase controls flux through the methylerythritol 4-phosphate pathway in Arabidopsis. *Plant Physiology,* 165**,** 1488-1504.

WU, T. D. & WATANABE, C. K. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics,* 21**,** 1859-1875.

XIA, Z., XU, H., ZHAI, J., LI, D., LUO, H., HE, C. & HUANG, X. 2011. RNA-Seq analysis and *de novo* transcriptome assembly of *Hevea brasiliensis*. *Plant Molecular Biology,* 77**,** 299.

XIAO, X., TANG, C., FANG, Y., YANG, M., ZHOU, B., QI, J. & ZHANG, Y. 2014. Structure and expression profile of the sucrose synthase gene family in the rubber tree: indicative of roles in stress response and sucrose utilization in the laticifers. *The FEBS Journal,* 281**,** 291-305.

XU, Q., HE, Y., YAN, X., ZHAO, S., ZHU, J. & WEI, C. 2018. Unraveling a crosstalk regulatory network of temporal aroma accumulation in tea plant (*Camellia sinensis*) leaves by integration of metabolomics and transcriptomics. *Environmental and Experimental Botany,* 149**,** 81-94.

YABE, S., IWATA, H. & JANNINK, J.-L. 2018. Impact of Mislabeling on Genomic Selection in Cassava Breeding. *Crop Science,* 58**,** 1470-1480.

YAMASHITA, S., MIZUNO, M., HAYASHI, H., YAMAGUCHI, H., MIYAGI-INOUE, Y., FUSHIHARA, K., KOYAMA, T., NAKAYAMA, T. & TAKAHASHI, S. 2018. Purification and characterization of small and large rubber particles from *Hevea brasiliensis*. *Bioscience, Biotechnology, and Biochemistry,* 82**,** 1011-1020.

YAMASHITA, S., YAMAGUCHI, H., WAKI, T., AOKI, Y., MIZUNO, M., YANBE, F., ISHII, T., FUNAKI, A., TOZAWA, Y., MIYAGI-INOUE, Y., FUSHIHARA, K., NAKAYAMA, T. & TAKAHASHI, S. 2016. Identification and reconstitution of the rubber biosynthetic machinery on rubber particles from *Hevea brasiliensis*. *Elife,* e19022.

YEANG, H. Y. 1986. Impedance of latex exudation by the bark excision wound during taping. *Journal of natural Rubber Research,* 1**,** 89-97.

YEANG, H. Y. 1998. Impact of biological variation on latex allergenicity. *Journal of Allergy and Clinical Immunology,* 101**,** 145-146.

YEANG, H. Y., ARIF, S. A. M., YUSOF, F. & SUNDERASAN, E. 2002. Allergenic proteins of natural rubber latex. *Methods,* 27**,** 32-45.

YEANG, H. Y., CHEONG, K. F., SUNDERASAN, E., HAMILTON, R. G. & CARDOSA, M. J. 1996. The 14.6 kd rubber elongation factor (Hev b 1) and 24 kd (Hev b 3) rubber particle proteins are recognised by IgE from patients with spina bifida and latex allergy. *Journal of Allergy and Clinical Immunology,* 98**,** 628-639.

YEANG, H. Y. & CHEVALLIER, M.-H. 1999. Range of *Hevea brasiliensis* pollen dispersal estimated by esterase isozyme markers. *Annals of Botany,* 84**,** 681-684.

YEANG, H. Y., YIP, E. & SAMSIDAR, H. 1995. Characterisation of zone 1 and zone 2 rubber particles in *Hevea brasiliensis* latex. *Journal of natural Rubber Research,* 10**,** 108-123.

YIP, E. 1990. Clonal characterisation of latex and rubber properties *Journal of natural Rubber Research,* 5**,** 52-80.

YOUNG, A. J. 1991. The photoprotective role of carotenoids in higher plants. *Physiologia Plantarum,* 83**,** 702-708.

YUSOF, F. & CHOW, K. S. 2003. *The biosynthesis of rubber,* Kuala Lumpur, Malaysian Rubber Board.

YUSOF, F., CHOW, K. S., WARD, M. A. & WALKER, J. M. 2000. A stimulator protein of rubber biosynthesis from *Hevea brasiliensis* latex. *Journal of Rubber Research,* 3**,** 232-247.

ZAKIA, A., GIBRAT, R., BRUGIDOU, C., PIERRE, T. & JEAN, D. A. 1992. Evidence for an amiloride-inhibited $Mg^{2+}/2H^{+}$ antiporter in lutoid (vacuolar) vesicles from latex of *Hevea brasiliensis*. *Plant Physiology,* 100**,** 255-260.

ZHANG, R. July 2017 2017. *RE: Exon size range for Arabidopsis for mapping optimisation.*

ZHANG, R., CALIXTO, CRISTIANE P. G., MARQUEZ, Y., VENHUIZEN, P., TZIOUTZIOU, N. A., GUO, W., SPENSLEY, M., ENTIZNE, J. C., LEWANDOWSKA, D., TEN HAVE, S., FREI DIT FREY, N., HIRT, H., JAMES, A. B., NIMMO, H. G., BARTA, A., KALYNA, M. & BROWN, JOHN W. S. 2017a. A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic Acids Research,* 45**,** 5061-5073.

ZHANG, R., CALIXTO, C. P. G., MARQUEZ, Y., VENHUIZEN, P., TZIOUTZIOU, N. A., GUO, W., SPENSLEY, M., FREI DIT FREY, N., HIRT, H., JAMES, A. B., NIMMO, H. G., BARTA, A., KALYNA, M. & BROWN, J. W. S. 2016. AtRTD2: A Reference Transcript Dataset for accurate quantification of alternative splicing and expression changes in Arabidopsis thaliana RNA-seq data. *bioRxiv*.

ZHANG, R., CALIXTO, C. P. G., TZIOUTZIOU, N. A., JAMES, A. B., SIMPSON, C. G., GUO, W., MARQUEZ, Y., KALYNA, M., PATRO, R., EYRAS, E., BARTA, A., NIMMO, H. G. & BROWN, J. W. S. 2015. AtRTD – a comprehensive reference transcript dataset resource for accurate quantification of transcript-specific expression in Arabidopsis thaliana. *New Phytologist,* 208**,** 96-101.

ZHANG, S.-J., WANG, C., YAN, S., FU, A., LUAN, X., LI, Y., SUNNY SHEN, Q., ZHONG, X., CHEN, J.-Y., WANG, X., CHIN-MING TAN, B., HE, A. & LI, C.-Y. 2017b. Isoform Evolution in Primates through Independent Combination of Alternative RNA Processing Events. *Molecular Biology and Evolution,* 34**,** 2453-2468.

ZHANG, X., GUO, T., XIANG, T., DONG, Y., ZHANG, J. & ZHANG, L. 2018. Quantitation of isoprenoids for natural rubber biosynthesis in natural rubber latex by liquid chromatography with tandem mass spectrometry. *Journal of Chromatography A,* 1558**,** 115-119.

ZHU, T. & WANG, X. 2000. Large-Scale Profiling of the Arabidopsis Transcriptome. *Plant Physiology,* 124**,** 1472-1476.